Englische Philologie

THE PHONETICS AND PHONOLOGY OF HONG KONG ENGLISH:

A STUDY OF FRICATIVES

Inaugural-Dissertation

zur Erlangung des Doktorgrades

der Philosophischen Fakultät

der Westfälischen Wilhelms-Universität

zu Münster (Westf.)

vorgelegt von

SIN YU BONNIE HO

aus Hongkong, China

2021

Tag der mündlichen Prüfung: 23.09.2021
Erstgutachter: Prof. Dr. Ulrike Gut
Zweitgutachter: Prof. Dr. Robert Fuchs

Acknowledgement

First and foremost, I would like to thank Prof. Dr. Ulrike Gut for all her guidance, support, and care throughout my PhD study. I would also like to thank Prof. Dr. Robert Fuchs and Prof. Dr. Dagmar Deuber for their valuable comments on the dissertation. I wish to thank Prof. Dr. Katerina Stathi, Prof. Dr. Dejan Matić, and the multitude of professors and colleagues from the Graduate School Empirical and Applied Linguistics and the English Department for teaching me a lot of things, which sharpened my research skills.

Also, as always, my appreciation to Prof. Dr. Hansen Edwards, who inspired me to study the phonetics and phonology of Hong Kong English. Special thanks to Dr. Florian Schiel and his colleagues as well as all the contributors for their continuous efforts in developing and improving Munich AUtomatic Segmentation (MAUS), which in my opinion, is one of the greatest research tools for phoneticians and phonologists.

With regards to data collection, I would like to thank Prof. Dr. Christophe Coupé and the colleagues from the Language Development Laboratory at the University of Hong Kong, and Prof. Dr. Stephen Politzer-Ahles and the colleagues from the Speech and Language Sciences Laboratory at the Hong Kong Polytechnic University for their assistance. I also thank DAAD-IPID4all and smartNETWORK International for funding the data collection in Hong Kong.

Many thanks to Dr. Effi Georgala from Nuance Communications, who is my manager at work but more like a mentor to me. Thank you for seeing the potential in me and always pushing me to my limit. Also, a big thank you to my PhD buddies: Laika Lo, Maggie Chak, and Adrian Derstroff. I would not have finished the dissertation without their support and company. Last but not least, I would like to express my greatest gratitude to my family in Hong Kong. They always stand by me no matter what.

Table of contents

List of tables

List of figures

Chapter 1

Introduction

Hong Kong English is one of the emerging or new varieties of English which has its own phonetic and phonological features. Previous studies on Hong Kong English suggested some characteristics such as a lack of long/short vowel contrasts and a lack of voiced/voiceless consonant contrasts (e.g. Bolton and Kwok, 1990; Hung, 2000; Setter et al., 2010). Hung (2000) also proposed an inventory of Hong Kong English with respect to vowels (see Figure 1.1) and consonants (see Table 1.1).

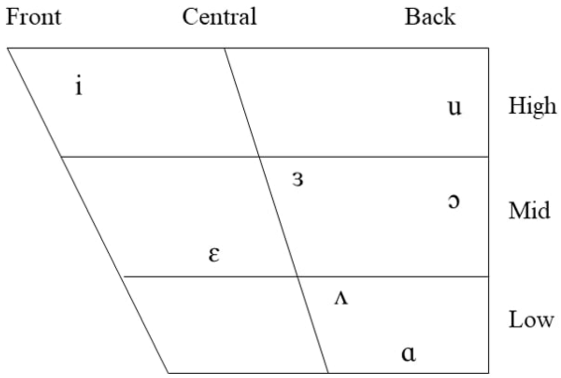

Figure 1.1 Vowel chart of Hong Kong English proposed by Hung (2000, p. 354)

Table 1.1 Consonant chart of Hong Kong English proposed by Hung (2000, p. 354)

	Bilabial	Labio-dental	(Inter-) dental	Alveolar	Palato/Post-alveolar	Palatal	Velar	Labio-velar	Glottal
Stop	p b			t d					
Affricate					ʧ ʤ				
Fricative		f	θ	s	ʃ				h
Lateral-approximant				l					
Approximant				r		j		w	
Nasal	m			n			ŋ		

Note: In the original study, /h/ was labelled as glottal approximant

As can be seen in Figure 1.1, the vowel inventory of Hong Kong English is smaller when compared to standard British or American English. Hung (2000) also conducted a formant analysis of the vowels and found that there are mergers of /iː/-/ɪ/, /e/-/æ/, /uː/-/ʊ/, and /ɔː/-/ɒ/ in Hong Kong English. Among consonants, fricatives seem to be the most interesting group because of its smaller inventory, as shown in Table 1.1, compared to standard British/American English. There is also a high variability in the realisations of fricatives in Hong Kong English as proposed by previous work (Hung, 2000; Hansen Edwards, 2019; Deterding, 2006).

Nevertheless, many predictions and hypotheses regarding Hong Kong English fricatives and their variants have not yet been tested systematically and quantitatively using a large dataset. One possible reason is that Hong Kong English alongside other new varieties of English is one of the low-resource language varieties and there are simply not enough tools accessible for linguists. By investigating the phonology of Hong Kong English, the computational methods which are desired for building speech and language technologies for low-resource languages and varieties can also be explored.

This study concerns the topics of phonetics and phonology, acoustic phonetics, variation of Hong Kong English fricatives, and computational methods for researching low-resource language varieties. Although some phonological theories and World Englishes models are touched on to facilitate the conceptualisation of Hong Kong English, this study concentrates on enriching the documentation of Hong Kong English phonology. Apart from the theoretical contributions to the field of Hong Kong English phonetics and phonology, the present study also attempts to apply the findings in automatic speech recognition (ASR) systems so as to improve the automatic phone recognition of Hong Kong English.

There are six major aims of this study, namely

(i) to determine the inventory of Hong Kong English fricatives and their variants

(ii) to investigate the acoustic characteristics of Hong Kong English fricatives and their variants

(iii) to examine the distribution and systematic variation of Hong Kong English fricatives

(iv) to generate phonological rules regarding Hong Kong English fricatives

(v) to establish a pipeline for speech data processing and acoustic analysis

(vi) to build an automatic speech recognition (ASR) model for Hong Kong English by adapting a state-of-the-art ASR system

This study is a large-scale quantitative study of Hong Kong English fricatives and can substantiate the phonological features of Hong Kong English reported in previous studies with statistical evidence. This study is one of the very few studies which examines the acoustic aspects of fricatives and their variants in Hong Kong English to provide more fine-grained details of the sound. Moreover, this study is the first study which attempts to i) build a classification model specific to Hong Kong English fricatives and their variants using neural networks, and ii) generate weighted pronunciation rules and apply them in an existing ASR model of English to improve phone recognition accuracy of Hong Kong English. The methods and findings of this study are not only beneficial to Hong Kong English researchers but all researchers who study new varieties of English and low-resource language varieties.

The structure of the rest of the dissertation is as follows.

Chapter 2 starts with a description of the sociolinguistic background of Hong Kong. It is followed by a brief overview of some popular World Englishes model, namely Kachru's Three Circles model and Schneider's Dynamic model. Then, different models on the conceptualisation of Hong Kong English are introduced. How variation in Hong Kong English phonology can be approached is discussed and it serves as a general framework of analysis for the present study. Finally, previous studies on Hong Kong English fricatives are discussed.

Chapter 3 is about the acoustics of English fricatives. First, the mechanical model of production of fricatives is introduced. It helps explain fricatives produced at different places of articulation. Then, the acoustic properties with respect to spectral, amplitudinal, and temporal properties of fricatives as well as the DCT coefficients, are presented. Since there is no previous research on the acoustic properties of Hong Kong English fricatives, the discussions are based on the comparison with standard American English fricatives. Finally, the methods of classifying fricatives in previous studies are discussed.

Chapter 4 gives an outline of the common procedures in a feature-based statistical ASR system. The Munich AUtomatic Segmentation System (MAUS), which is a forced alignment tool for phonetic segmentation and labelling, is introduced. The syntax of the MAUS pronunciation rule set is also delineated.

Chapter 5 lists the research questions and predictions based on the findings of previous studies and the discussions in previous chapters.

Chapter 6 details the method employed in this study with participants, materials, and data collection procedures. In terms of data processing, a pipeline was created in order to automate several steps in acoustic analysis. Acoustic analysis using the subset of the word list data was conducted and auditory analysis was conducted using the full set of word list data. Classification models of place of articulation, voicing, and phones were trained using convolutional neural networks. The statistical models used in this study and the architecture of the convolution neural networks are described in this chapter.

Chapter 7 presents the results of the acoustic analysis of fricatives and their variants. The smoothed spectra using discrete cosine transform (DCT) coefficients are plotted to provide visual evidence. The findings are discussed with respect to the research questions stated in Chapter 5.

Chapter 8 reports the classification performances of Hong Kong English fricatives and their variants with respect to place of articulation, voicing, and phone symbols. Results of the error analysis of the classifications are also discussed.

Chapter 9 reports the findings from the auditory analysis of the full word list dataset. The findings are discussed with respect to the research questions stated in Chapter 5. An inventory of Hong Kong English fricatives is also proposed. Potential phonological rules of Hong Kong English fricatives derived from the findings of this study are presented.

Chapter 10 discusses applicability of the results of this study in automatic speech recognition (ASR) systems. Results of the adapted acoustic and language model of (standard) British English in MAUS with Hong Kong English fricatives pronunciation rules are presented. The possibility of applying phonological features in ASR systems to improve recognition performance is also discussed.

Chapter 11 concludes the present study. The contributions of this study and future research directions are presented.

Chapter 2

Hong Kong English phonology

2.1 Sociolinguistic background of Hong Kong

Hong Kong is a city located in the southern coast of China, adjacent to the Pearl River Delta (Setter et al., 2010). The official name of Hong Kong is Hong Kong Special Administrative Region of the People's Republic of China (HKSAR). Although Hong Kong is a city in China, it is very different from other cities in mainland China in various ways such as law, currency, official languages, and education system. Over the past centuries, the political, economic, and social environments, as well as the education system and language policy in Hong Kong, have been constantly changing. Hence, Hong Kong is no-doubt an exciting place for conducting sociolinguistic research. Hong Kong experienced several changes of sovereignty: i) ceded to Britain after the Opium War in 1842, ii) occupied by Japan during 1941-1945, iii) resumed to be a British colony since 1945, and iv) handed over to China in 1997. That is to say, Hong Kong was a British colony for more than a hundred years. During the British colonial period, Hong Kong had transformed from an entrepôt into an industrial and manufacturing center in the 1950s and evolved into an international financial center in the 1980s. Since the 2010s, there have been various protests and movements against the Central and local government, such as the protests against the National Education in 2011-2012, the Umbrella Movement in 2014, and Anti-Extradition Law Amendment Bill Movement in 2019-2020. These sovereignty, economic, and political changes not only affect the language policies but also the language use, attitudes, and identity (Hansen Edwards, 2018).

The mother tongue or first language of the majority of Hong Kong people is Cantonese, which is a variety of Chinese. With regards to English, English was first introduced during the British colonial period. Bolton and Kwok (1990) described the linguistic situation back then as "a case of societal bilingualism in which two largely monolingual communities co-exist, with a small group of bilingual Cantonese functioning as 'linguistic middlemen'" (p. 148). Luke and Richards (1982) referred to the linguistic situation as diglossia without bilingualism. During the colonial period, English was considered a high language, and was mainly used in government bodies like courts and Legislative Council, official bills and documents, and at university, whereas Cantonese was considered a low language and was mainly used in private domains such as home, among friends, and for intranational communication (Groves, 2011). Nevertheless, English has been widely spread by being a compulsory subject in the primary and secondary school curriculum. There are also many schools which use English as the medium of Instruction (EMI). As a result, there is an increasing population who are able to speak English. Due to the handover to China, the Biliterate and Trilingual language policy, meaning biliterate in written Chinese and English, and trilingual in Cantonese, English, and Mandarin, was introduced in 1997. Since then, Mandarin has been introduced as a compulsory subject in primary and secondary schools.

According to the population census in 2016, Hong Kong population reached 7.3 million (Census and Statistics Department, 2016). The literacy rate was 96% and almost one-third of the population received post-secondary education. Cantonese remained the most common language with 88.9% of the population speaking Cantonese at home. 53.2% of the population were able to speak English and 48.6% of the population were able to speak Mandarin. What is particularly interesting with respect to the present study is the linguistic profile of the population who received education in the post-colonial period. Over 95% of full-time students aged between 6 and 24 (as of 2016) were bilingual in Cantonese and English, and over 60% of full-time students aged between 6 and 24 were trilingual in Cantonese, English, and Mandarin (Census and Statistics Department, 2016). For this specific population, it is surprising to see that the rate of trilingualism was only around 60% since both English and Mandarin were taught in primary and secondary schools. One speculation is that Cantonese and English are the dominant languages in everyday life while Mandarin is seldom used except when communicating with people from mainland China. Another speculation is that the discontent over the Central and local government alongside the political protests and movements in the 2010s is also reflected in the language use, attitudes, and identity of this population.

2.2 Conceptualisation of Hong Kong English

Hong Kong English generally refers to the English spoken by people from Hong Kong. There have always been debates about whether Hong Kong English exists (Luke and Richards, 1982) or whether it can be regarded as a new or emerging variety of English (Groves, 2009). The status and usage of English as well as the developmental processes of a new variety of English can be described by different frameworks and models. Although there is no single model which can perfectly define a language variety or fully capture how a language variety is developed, looking into different models helps better understand the complexity of the issue and how Hong Kong English can be conceptualised. In this section, Kachru's classic Three Circles model, Schneider's Dynamic model, and Buschfeld and Kautzsch's Extra- and Intra-territorial Forces model are briefly reviewed. Previous studies on how to model and conceptualise Hong Kong English (Bolton and Kwok, 1990; Hung, 2000; Q. Zhang, 2013) are also examined.

Kachru's Three Circles model classifies countries into Inner Circle, Outer Circle, and Expanding Circle based on the spread and usage of English (Kachru, 1985). Inner Circle refers to traditional English-speaking countries such as the USA, the UK, Canada, Australia, and New Zealand in which English is used as a first language and in almost all domains (both high and low). Outer Circle concerns countries where English is spread and used as a second language (ESL), mostly through colonisation. Examples of Outer Circle countries include Nigeria, India, and Singapore. English is usually an official language in Outer Circle countries and "has undergone some acculturation and nativisation" (Groves, 2009, p. 56). Expanding Circle countries include Japan, Korea, and China, in which English is used as a foreign language (EFL). In the EFL classrooms, often Inner Circle Englishes are taught as the norm.

With regards to the status and usage of English in Hong Kong, it is true that Hong Kong was a colony of Britain, and English is mainly used as a second language. Hence, Hong Kong can be considered as one of the Outer Circle countries. Nevertheless, when compared to other Outer Circle countries like Singapore and Malaysia in which English serves as a lingua franca between ethnically and linguistically diverse groups (Bolton and Kwok, 1990), English in Hong Kong is mostly spoken in school or work settings but seldom in private settings. Instead, English is heavily mixed in Cantonese speech (Bolton and Kwok, 1990). It is because the population is relatively homogeneous. More than 90% of the population is ethnic Chinese and speak Cantonese as their primary language (Groves, 2009). Therefore, Kachru's Three Circles model lacks the flexibility to describe English in Hong Kong as the sociolinguistic situations are more complex and dynamic (Schneider, 2007). Another critique of the model is that the Three Circles model fails to address the 'pluricentrality' of English nowadays for the norms should no longer be determined by Inner Circle countries (Schneider, 2003).

Schneider (2007) proposed that post-colonial Englishes generally undergo similar consecutive developmental stages, namely i) foundation, ii) exonormative stabilisation, iii) nativisation, iv) endonormative stabilisation, and v) differentiation. In the foundation stage, English is brought to a new territory by a dominant group of settlers. A clear dichotomy between 'other' and 'us' exists, meaning there are two separate social groups: the settler strand and the colonised/indigenous strand. The two groups are distinct from each other in terms of not only ethnicity but also the languages they use and their social networks. In the exonormative stabilisation stage, English is officially established as the language of administration, education, the legal system, etc. Segregational elitism can be found with the settlers dominating the higher governing positions. There is also a form of language elitism. Standard British or American English is held as the exonormative model of English and is generally preferred. Also, if people from the indigenous strand can speak or use English, they are considered elite and have a higher social status. In this stage, linguistic transfer on the levels of phonology and structure starts to occur due to the inevitable language contact of English and the indigenous group's vernacular. In the nativisation stage, a marked local accent of English is shown among the indigenous strand speakers as a result of language contact, and code-mixing of English and the vernacular is observed. Debates and discussions on the legitimacy of the endonormative form of English begin to emerge. In the endonormative stabilisation stage, the local linguistic norm is accepted also in formal contexts and high domains. A linguistic independency and political independency (Groves, 2009) is achieved, meaning the endonormative variety of English can claim its independency from the exonormative Englishes in all domains and achieve "a cultural self-reliance" (Schneider, 2007, p. 48) with the full acceptance from the indigenous strand. In this stage, "X English" substitutes the label of "English in X", which signals "different conceptualizations of the status of the language" (Schneider, 2007, p. 50). According to Schneider (2007), Singapore and Jamaica are examples of this stage. Finally, in the differentiation stage, internal and individual differentiation begins to bloom. Individuals from the indigenous group can exert their "personal predilections" and create their own forms of English within the "new nation" (Schneider, 2007, p. 53), which represents the new social identities (Groves, 2009).

One emphasis of Schneider's Dynamic model is that the development of a new variety is not static, and it is not solely dependent on the linguistic effects but also on the sociopolitical background, identity construction, and socialinguistic conditions. Over the past few decades, Hong Kong English scholars have been probing into these factors by conducting attitudinal and language identity studies (e.g. Hansen Edwards, 2015; Hansen Edwards, 2018; Q. Zhang, 2013), and documenting the phonological, lexical, syntactic, and discoursal features of Hong Kong English (e.g. Bolton, 2002; Setter et al., 2010; Deterding et al., 2008). The claim by Luke and Richards (1982) that "there is no societal basis for 'indigenisation' or 'nativisation' of English in Hong Kong" has been refuted. Schneider (2007) has categorised Hong Kong to the nativisation stage. Recent studies have found that Hong Kong English may have entered the endonormative stabilisation stage in the Dynamic model, as the endonormative linguistic features are relatively stabilised (Setter et al., 2010) and there is growing acceptance of the Hong Kong English accent by viewing it as a local identity marker (Hansen Edwards, 2015). At the same time, Hong Kong English is still not preferable in formal contexts (e.g. news broadcast and business meeting) when compared to Inner Circle varieties of English (Jim Y.H. Chan, 2016), suggesting that Hong Kong English has not fully reached the endonormative stabilisation stage in the Dynamic model.

Although the development of Hong Kong English is, to a large extent, influenced by the settler (British colonial) strand, other external and internal factors should also be taken into consideration. Buschfeld and Kautzsch (2017) suggested a model of extra- and intra-territorial forces, in which five major subcategories of forces were proposed, namely i) colonisation or attitudes towards colonisation, ii) language policies, iii) globalisation, iv) foreign policies, and v) the sociodemographic background of a country. Globalisation as an extra- and intra-territorial force refers to the "linguistic and also cultural influences coming from the Internet, US popular culture, and modern media as well as trading relations between countries...[and] to whether and to what extent they accept or even admit these facets of globalisation" (Buschfeld and Kautzsch, 2017, p. 11). This is particularly relevant to Hong Kong as Hong Kong is perceived as one of the international cities. Deterding et al. (2008) also noted that while the English accent of many people from Hong Kong was based on British English, there were clear American influences in their speech data.

When it comes to the modelling of a new variety of English, what recent studies have advocated is that instead of a static handling, the developmental processes of a variety should be conceptualised as a continuum (Schneider, 2007; Buschfeld and Kautzsch, 2017). In the case of Hong Kong, regardless of which stage Hong Kong English is in, there is always variability, especially when it comes to accent (Hung, 2000). Therefore, the dynamic handling can be applied to the phonological system of Hong Kong English. Bolton and Kwok (1990) is one of the very first studies that proposed a dynamic model of Hong Kong English accent (see Figure 2.1). According to the model, speakers generally share some localised features of Hong Kong English (represented in bold triangle) and the features are subject to "a good deal of variation" (Bolton and Kwok, 1990, p. 166). Apart from the Hong Kong English features, the speakers may also use "phonological forms that approximate to the reference systems of [British] English and American English" (Bolton and Kwok, 1990, p. 166) (represented in triangles with solid lines at two ends). What Bolton and Kwok (1990) observed in their speech data was that instead of an either-or decision between different forms, there was clustering of items of Hong Kong English features and (standard) British English or American English forms (represented in triangles with dashed lines). This model contrasts with previous postulates by Luke and Richards (1982), which assumed that speakers (or "learners" in the original study) would unanimously move towards the norm of a British or American English accent when their English proficiency increased.

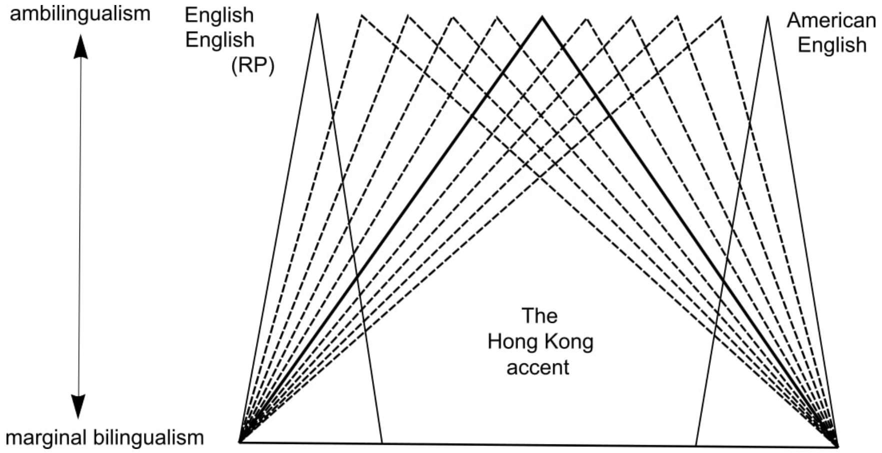

Figure 2.1 The dynamics of a Hong Kong accent (Bolton and Kwok, 1990, p. 166)

Similarly, Hung (2000) studied the phonology of Hong Kong English and stated that "[the] internalised phonological system of an individual speaker of [Hong Kong English] is, like any interlanguage system, dynamic and evolving rather than static" (p. 339). He also proposed a continuum with an idealised Hong Kong English phonology which consisted of all endonormative phonological features at one end, and the standard American or British English phonology at the other end (Hung, 2000). He acknowledged that speakers of Hong Kong English might spread across the continuum and differ in how many phonological features of the idealised Hong Kong English were present in their speech.

In line with Hung (2000), Q. Zhang (2013) who studied attitudes towards Hong Kong English accent conceptualised Hong Kong English phonology as a continuum with broad accent (HKbr) at one end and educated accent (HKed) on the other. The broad accent was similar to the "mesolect...marked by a high frequency of [Hong Kong English] features" and the educated accent was similar to the "acrolect" spoken by people who were close to the "exonormativity of the American or British English accent but with localised features remained" (Q. Zhang, 2013, p. 10). Figure 2.2 is an illustration of the continuum of Hong Kong English phonology.

The present study adapts the proposed model in Q. Zhang (2013) and conceptualises Hong Kong English phonology as a continuum. This study also expects a certain variability in the system. It would be interesting to review how previous studies approached variability (see Section 2.3) and to know if the variation or the use of variants are due to internal or linguistic factors (e.g. syllable position and stress) or due to external or non-linguistic factors (e.g. gender, age, and proficiency) (see Section 2.4).

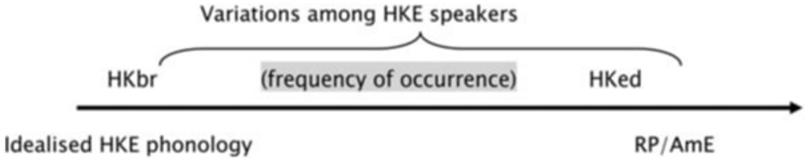

Figure 2.2 The continuum of Hong Kong English (HKE) phonology (Q. Zhang, 2013, p. 113)

2.3 Towards variation and Hong Kong English phonology

Language variation can be studied differently based on three different interpretations. The first line of research into new varieties of English focuses on identifying specific localised features, and how they are different from the phonological systems of Inner Circle Englishes. Schneider (2007) explained that such a kind of variation can be identified as "transfer phenomena from the phonology of indigenous languages" (p. 44). In the context of Hong Kong, although Cantonese is the indigenous language, due to the trilingual language policy, Mandarin also starts to play an important role in the linguistic repertoire of speakers who were born in the post-colonial period. Therefore, the phonological system of Cantonese and/or Mandarin is often referenced and compared (such as in A. Y. Chan and D. C. Li, 2000; Setter et al., 2010; Hansen Edwards, 2019; Hung, 2000). Based on this interpretation of variation, previous studies on Hong Kong English phonology also examined the uniqueness of the localised features by comparing them with other new varieties of English in South-East Asia, such as Singapore English, Malaysian English, Mainland Chinese English, and Vietnam English (Hansen Edwards, 2016; Deterding et al., 2008; Hung, 2000).

The second interpretation of variation is an extension of the first interpretation and examines the internal or linguistic system such as under which conditions variation in the phonetic realisations occurs. Previous studies on Hong Kong English phonology have attempted to investigate a vast number of linguistic factors which influence the realisation of fricatives, such as syllable position, word position, stress, singleton or consonant cluster, preceding and following phonetic environments, morphological conditioning, and lexical frequency (Hansen Edwards, 2016; Hung, 2000; Deterding et al., 2008).

The final interpretation of variation is also an extension of the first interpretation and it looks into the inter-speaker levels. Studies along this line examine the variation between different speaker groups, usually based on a number of social variables such as age, gender, status, and education level (Hansen Edwards, 2019). What Schneider (2007) generally observed in many post-colonial communities is "a range of sociolinguistic variation...with proximity to native speakers' pronunciation forms increasing in correlation with status, education, and frequency of interaction with them" (p. 44). It should be noted that apart from inter-speaker variation, there is also variation in the intra-speaker level as studies on new or emergent varieties of English also found variation or inconsistencies within a single speaker

(Deterding et al., 2008). Schneider (2003) predicted that in the course of time, the amount of such kind of intra-speaker variability will be reduced.

To sum up, studying variation using the above interpretations helps provide a more holistic understanding of the phonological system of Hong Kong English. While studying the inter-speaker variation is equally important, the present study focuses on inspecting the internal linguistic variation in the Hong Kong English phonology. To probe this issue, this study adopts the three main components listed in Hung (2000, p. 338) as a general framework of analysis and discussion, namely:

1. An inventory of phonemes, or sound segments which contrast with each other.

2. Systematic variation in the phonetic realisations of these phonemes, i.e. alternation.

3. The distribution of individual segments in relation to other segments.

In the next section, findings of the phonological features from previous studies on Hong Kong English are mainly discussed with respect to these three components.

2.4 Previous studies on Hong Kong English fricatives

Bolton and Kwok (1990) is one of the earliest studies which described the phonological features of Hong Kong English. The aim of their study was to illustrate as many of the localised features as possible based on impressionistic or intuitive judgement. They presented findings from an interview with a university student, whom they referred to as a "mid-range" speaker of Hong Kong English. Received Pronunciation (RP) was adopted as a reference for comparison. The consonant features were classified by three main types of phonological processes:

(i) Deletion
(ii) Substitution
(iii) Devoicing of voiced consonants

Generally, it was found that the realisation of consonants was influenced by the phonology of Cantonese, to a large extent. Table 2.1 is an overview of English and Cantonese consonants extracted from A. Y. Chan and D. C. Li (2000, p. 68). Regarding dental fricatives, Bolton and Kwok (1990) found that /θ/ was substituted by [f]. This phenomenon is also called TH-fronting. /ð/ was replaced by [d] in word-initial position. This phenomenon is also called TH-stopping. /ð/ was also substituted by [v̥] (presumably [f]) in word-final position. Regarding labiodental fricatives, /v/ was replaced by [w]. The process can also be called gliding. Postalveolar fricative /ʃ/ was substituted by [s] since there was no /ʃ/ in Cantonese but /s/ (see Table 2.1). In terms of deletion, only non-release of plosives in the word-final position was noted but the deletion of fricatives was not reported. Bolton and Kwok (1990) explained that it might be due to the unreleased /p, t, k/ in checked syllables in the Cantonese phonology. There was also devoicing of voiced consonants for fricatives, as well as for plosives. Nevertheless, findings from their study have limited generalisability. Not many conclusions can be drawn based on the speech of only one speaker and without any frequency count of each realisation. There also lacked an in-depth phonological analysis of the realisation of fricatives.

Table 2.1 An overview of English and Cantonese consonants adapted from A. Y. Chan and D. C. Li (2000, p. 68)

Manner of		Place of articulation								
articulation		Bilabial	Labiodental	Dental	Alveolar	Postalveolar	Palatal	Velar	Labiovelar	Glottal
Plosive	E	p b			t d			k g		
	C	p b			t d			k g	kʷ gʷ	
Fricatives	E		f v	θ ð	s z	ʃ ʒ				h
	C		f		s					h
Affricates	E					ʧ ʤ				
	C				ts dz					
Nasals	E	m			n			ŋ		
	C	m			n			ŋ		
Lateral	E				l					
	C				l					
Approximants	E	w				ɹ	j			
	C	w					j			

E: English C: Cantonese

Hung (2000) conducted a comprehensive study of Hong Kong English phonology, in which variation was systematically studied. He collected the speech data of 15 university students using a word list with minimal or similar-sounding pairs. The reason why he employed a word list instead of a reading passage was to limit the effects from other linguistic variables such as "differences in sentence structure, word and sentence stress, intonation" (Hung, 2000, p. 339). He also preferred a word list to spontaneous speech because it was challenging to elicit all possible "phonemic contrasts and allophonic variation in a speaker's phonological system from this sort of random data" (Hung, 2000, p. 339). What he found was that there was no evidence of a voiced/voiceless contrast of fricatives for most of the speakers. Specifically, there were no tokens of [z] and [ʒ] in all word-initial, -medial, and -final position. They were substituted by [s] and [ʃ] respectively. /v/ also did not exist in Hong Kong English, and it was realised as [f] and [w]. It is surprising to see /v/ realised as [w] in word-medial and intervocalic position regardless of stress pattern (e.g. *province* [ˈpɹoʊwins] and *provincial* [pɹoʊˈwinʃəl]) in his data. It is contrary to previous proposition by Edge (1991) that /v/ was only realised as devoiced [v̥] (or [f]) even in intervocalic position. Regarding dental fricatives, more than half of the speakers produced instances of [θ], whereas for other speakers, they produced [f] instead (TH-fronting), and in all phonetic environments. It showed that for these speakers, /θ/ was not part of their phonological system. As for /ð/, there was no evidence of it and the tokens were pronounced as [d] in word-initial or intervocalic position (TH-stopping), and as [θ] in word-final position.

Although it is generally noted that voiced consonants are devoiced in Hong Kong English, the behaviour of fricatives is different from that of plosives. For example, Hung (2000) found that the voiced/voiceless contrast of plosives remained. He suggested that the voiced plosives were distinguishable from the voiceless plosives by aspiration and voice onset time in word-initial position. Voiced plosives were non-aspirated and had shorter voice onset time. Since the difference was phonemic, he continued to adopt the conventional voiced/voiceless labels for plosives. As for fricatives, there were no distinctions between the voiced and voiceless fricatives. Therefore, he proposed the consonant chart regarding plosives and fricatives as shown in Table 2.2.

Table 2.2 Inventory of Hong Kong English plosives and fricatives proposed by Hung (2000, p. 355)

	Bilabial	Labiodental	(Inter-)dental	Alveolar	Palato-/Post-alveolar	Velar
Stop/Plosive	p b			t d		k g
Fricative		f	θ	s	ʃ	

Hung (2000) provided a solid foundation for studying the variation in Hong Kong English phonology and suggested some linguistic factors (mainly word position). Nevertheless, it was not known how the realisation would be in connected speech.

Deterding et al. (2008) examined the pronunciations of 15 Hong Kong university students using interview data. The measurements and frequency counts of the features were reported. The findings were also compared with the features of other varieties of English in Southeast Asia. Regarding dental fricatives, /θ/ (or voiceless TH used in the original study) appeared 61 times in total (42 times in word-initial position, 10 times in word-medial position, and 9 times in word-final position). In word-initial position, the majority (around 60%) were pronounced as [θ], around one-third of the tokens as [f], and one token as [t]; in word-medial position, [θ] is most common (60%), two occurrences as [t] and as omission respectively; and in word-final position, [f] was most common (around two-third). It was noted that for some speakers, the use of [f] was consistent in all positions, meaning only /f/ but not /θ/ was in their phonological systems.

The realisation of /θ/ as [f] in word-initial position was different from the realisation as [s] in (mainland) China English (Deterding, 2006; Rau et al., 2009). One possible explanations is that the acoustics of Cantonese /f/ is more similar to English /θ/ than the acoustics of Cantonese /s/ to English /θ/ (Deterding et al., 2008; Hansen Edwards, 2019). In fact, there are more acoustic similarities between /f/ and /θ/ (which are both non-sibilant fricatives) than /s/ and /θ/ in Inner Circle varieties of English (Jongman et al., 2000). The realisation of /θ/ as [t] might be an "exception rather than the norm" (Deterding et al., 2008, p. 155), unlike in Singapore English, in which [t] occurred more frequently (Deterding, 2007). As for /ð/ (or voiced TH), /ð/ was generally pronounced as [d] in the word-initial position. Since there were no instances of /ð/ in word-medial and word-final position, not much can be concluded.

Although Deterding et al. (2008) touched on many aspects of Hong Kong English such as vowels, consonants, rhythm, and sentence stress, with respect to fricatives, only the phenomenon of TH-fronting and TH-stopping was reported and discussed in the original paper. The problems in Deterding et al.'s study were well-pinpointed by Hung (2000). The analyses and discussions were limited due to the scarce data, which is one of the pitfalls of using spontaneous data. The scarce data was also due to small sample size, as there were only 15 speakers and the averaged interview duration per speaker was only 160s. The authors also acknowledged this limitation: "[w]e would need more extensive data from each speaker in our Hong Kong data to check whether Hong Kong speakers really are mostly consistent in their usage" (Deterding et al., 2008, p. 155).

There are other studies which attempt to delineate the phonology of Hong Kong English in a broad manner, meaning they looked into different aspects of Hong Kong English phonology (see Setter et al., 2010; Sewell and J. Chan, 2010). Together, previous studies seem to suggest some systematic variation regarding fricatives, which can be statistically tested. Hansen Edwards (2019) is one of the very few studies, if not the only study, which inspected Hong Kong English phonology using VARBRUL analysis. VARBRUL is basically an implementation of multiple logistic regression of variable data and is developed for sociolinguistic and language variation studies (Bayley, 2002; Paolillo, 2002). Hansen Edwards (2019) examined the realisation of voiceless dental fricatives /θ/ with respect to both linguistic factors (syllable position, linguistic environment, and stress) and social factors (gender, medium of instruction, and proficiency) using interview and reading data from 44 university students. Results showed that significantly more participants with advanced proficiency had /θ/ in their phonological system than participants with low intermediate proficiency. TH variation also occurred more often in low intermediate participants. It implied that TH variation might be a developmental phenomenon (Hansen Edwards, 2019). Among the 44 participants, 24 of them demonstrated TH variation and their data were further analysed (n=1680). 67% of TH were realised as [θ], 29% as [f], and 5% as [s]. VARBRUL analysis demonstrated that stress and word position were not significant predictors. Regarding the social factors, gender, task style, and educational background were also not significant. Nevertheless, for the realisation as [f], lexical category (numerals), following linguistic environment (/r/), syllable position (onset), and preceding labial (e.g. /m/) in the same word were found to be significant. As for the realisation as [s], following linguistic environment (vowel and /r/), preceding labial in the same word, proficiency (advanced), and preceding linguistic environment (preceding obstruent) were found to be significant factors.

That TH variation is a developmental phenomenon in Hong Kong English is not surprising as research on both first language and second language acquisition of English also found that the realisation of /θ/ as [f] was more common in early stages of acquisition (Hansen Edwards, 2019). It is also in line with the conceptualisation of Hong Kong English as a continuum that there are more occurrences of localised features when moving towards the lower end of the continuum. There was a small amount of instances of [s], suggesting that this variation might be due to the influence of Mandarin (Deterding, 2006; Rau et al., 2009). While the realisation of /θ/ as [f] or [s] were governed by several linguistic and social factors, [f] and [s] were not allophonic variation in Hong Kong English "but rather emerge[d] at different stages in the acquisition of TH" (Hansen Edwards, 2019, p. 26). Nevertheless, it is hard to explain why advanced participants were more likely to produce [s] than high intermediate and low intermediate participants, given that all participants in the study received education of English and Mandarin at school. It was postulated that "as speakers of English in Hong Kong gain higher levels of proficiency in Mandarin Chinese, features of Mandarin Chinese will emerge in Hong Kong English" (Hansen Edwards, 2019, p. 26). Nevertheless, being proficient in English does not necessarily imply that the participant was also proficient in Mandarin. Such an assumption requires more support.

The lexical analysis in Hansen Edwards's study was a discussion of findings with respect to the linguistic factors and per lexical items. Hansen Edwards (2019) demonstrated the difficulty in interpreting the results such as why the realisation as [f] occurred more frequently in the numerals *third, thirty, thirteen* and [f] was favoured more in *think* than in *thinking.* Rau et al. (2009), who studied the variation of English voiceless dental fricatives by Chinese speakers, found that lexical frequency had an effect on the realisation of /θ/. A proposition can be devised that speakers are more familiar with high frequency words than the low frequency words, and different lexical familiarity will affect the realisation of fricatives. Nevertheless, the underlying assumption is that the dataset is large enough to cover various English fricatives. Similar to the study by Deterding et al. (2008), there were many scarce data in the study by Hansen Edwards (2019), in which many lexical items only appeared once or just a few times.

To conclude, most studies on Hong Kong English phonology adopted a more impressionistic or exploratory approach without providing much quantitative support. Nevertheless, their findings suggest that there seems to be systematic variation with respect to fricatives in the phonology of Hong Kong English. For example, the variation of /w/ and /ð/ is related by word position; the realisation of /θ/ as [θ] remains the majority and the variation is related to syllable position, as well as the preceding and following phonetic environment. English proficiency also affects the realisation of /θ/. The substitution for /z/ and /ʒ/ occurs in all contexts. Table 2.3 is a summary of the findings and discussions from previous studies.

Table 2.3 Features and potential linguistic factors/hypotheses of fricatives suggested by previous studies on Hong Kong English Phonology

Feature	Description	Potential factors/hypotheses
/v/ substitution	/v/ is substituted by [f] or [w]	There is no phoneme /v/ [w]: word-initial position and intervocalic position [f]: word-final position
/θ/ variation	/θ/ is realised as either [θ] or [f]/[s]	[f]/[s]: syllable onset position, followed by /r/, preceding labial in the same word, English proficiency, lexical item
/ð/ substitution	/ð/ is substituted by either [θ] or [d]	There is no phoneme /ð/ [d]: word-initial or intervocalic position [θ]: word-final position
/ʃ/ substitution	/ʃ/ is substituted by [s]	Substitution takes place in all contexts
/z/ devoicing/	/z/ is substituted by [s]	There is no phoneme /z/ Substitution takes place in all contexts.
/ʒ/ devoicing/	/ʒ/ is substituted by [ʃ]	There is no phoneme /ʒ/ Substitution takes place in all contexts.

(Bolton and Kwok, 1990; A. Y. Chan and D. C. Li, 2000; Deterding et al., 2008; Hansen Edwards, 2019; Hung, 2000; Sewell and J. Chan, 2010)

Although Hansen Edwards (2019) proposed some social factors, most studies suggest that the realisation of fricatives is governed by linguistic factors. Therefore, this study focuses on examining the variation with respect to linguistic factors. The present study aims to evaluate the realisation of fricatives (mainly /v/, /θ/, /ð/, /ʃ/, /z/, and /ʒ/ as shown in Table 2.3) with respect to different linguistic factors (mainly syllable position and preceding labial in the same word) using a quantitative approach.

There are several takeaways from the previous studies, which help devise the method for this study, namely:

(i) to test the potential factors/hypotheses listed in Table 2.3 requires a large corpus of data so that statistical inferences can be made,

(ii) in terms of the materials used to elicit speech data, although it is important to examine the production in spontaneous speech data, what previous studies demonstrated are challenges to collect utterances in different phonetic environments and with sufficient instances from spontaneous speech data,

(iii) it is difficult to eliminate the lexical frequency and lexical familiarity effect when using real-word data

Therefore, this study used a word list comprising of different pseudo-words embedded in a carrier phrase. By doing so, the production of fricatives in various phonetic environments can be elicited systematically. In addition, a reading passage was employed to collect real-word data to complement the findings from the word list. Moreover, almost all previous studies conducted only an impressionistic and auditory analysis of fricatives. The fine-grained details of fricatives, especially of the non-sibilant fricatives, could not be captured. Therefore, this study also conducts acoustic analysis of the fricatives. Details of the method and materials are presented in Chapter 6.

Chapter 3

Acoustics of English fricatives

3.1 The production of fricatives

Fricatives are produced when air passes through a narrow constriction in the vocal tract. Turbulence, which is irregular and random air molecule motion, is created during the production of fricatives. Turbulence results in an aperiodic acoustic wave form, which is different from the complex periodic wave form generated by the vibrating vocal folds during voicing. The occurrence of turbulence is dependent on two major factors, namely the size of the vocal tract and the volume velocity of airflow (Shadle, 1985). It was found that when the vocal tract is constricted to around 10 mm^2, turbulence noise can be produced (assuming normal volume velocity of airflow) (Johnson, 2011). Mechanical models demonstrated that when the volume velocity of airflow increases, the turbulence noise also increases and at higher frequencies and amplitude (Shadle, 1985). Apart from the constriction, noise is generated when the jet of turbulence hits a downstream obstacle (see Figure 3.1), which is usually an articulator in the oral cavity, such as the teeth. Even though the production of fricatives involves a constriction and/or an obstacle, it is not completely blocking the airstream. Hence, the speech sound produced is a continuant, which is contrary to an occlusive such as a stop.

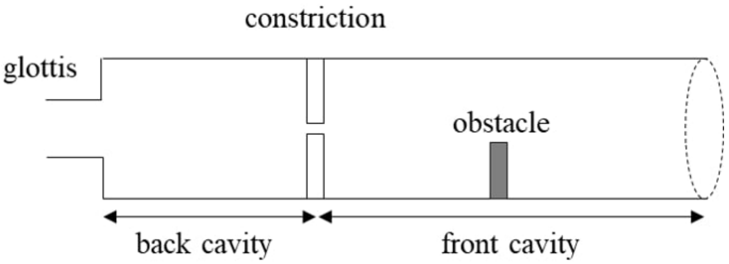

Figure 3.1 Tube model of an obstacle fricative adapted from Johnson (2011, p. 155)

Most English fricatives involve an obstacle, except for /h/, for which the fricative constriction is located at the glottis and the shape of the constriction is almost parallel to the airflow. For other articulators, they act as an obstacle which is more perpendicular-like to the airflow. In terms of acoustics, such kind of obstacles produce "periodic vortices" which "contribute high-frequency components to the spectrum" of that sound (Johnson, 2011, p.156). At the same time, there may be a dampening effect caused by the obstacle that affects the frequency responses of the cavity.

Place of articulation

English fricatives (excluding /h/) can be categorised into four places of articulation, namely labiodental, (inter-)dental, alveolar, and postalveolar (which is sometimes used interchangeably with palato-alveolar).

/f/ and /v/ are labiodental fricatives, meaning that the lower lip and the upper teeth form a constriction, while the upper lip acts as an obstacle, and the turbulence is made at the lips. Since /f/ and /v/ are produced at the lips, which is probably the furthest part of the vocal tract, there is not much room left for the front cavity to filter[1] the sound. Hence, a more diffuse and flat spectrum is expected for labiodental fricatives.

/θ/ and /ð/ are dental fricatives. They are produced by placing the tip of the tongue at the back of the upper teeth. Dental fricatives are also called interdental fricatives because they can be produced by placing the tongue between the upper and lower teeth as well. Compared with other fricatives, dental fricatives sound weaker because the noise source of dental fricatives is close to the constriction area, as illustrated in Figure 3.2. It encounters higher "acoustic impedance" than the noise source located further to the constriction, and is less effective in "exciting the front cavity" (Zhao, 2010, p. 128). Moreover, there is no obstacle in the front cavity to create additional turbulence.

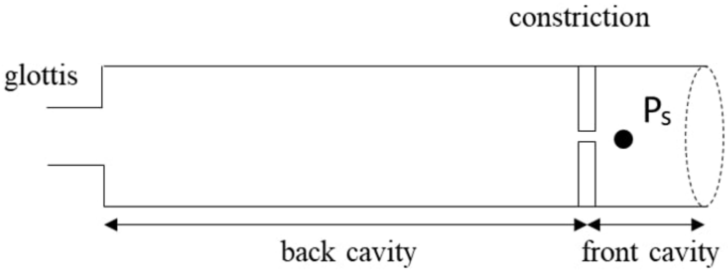

Figure 3.2 Tube approximation of a dental fricative adapted from Zhao (2007, p. 20) which shows that the turbulent noise source, denoted as P_s, is close to the constriction.

/s/ and /z/ are alveolar fricatives, for which the constriction is formed by placing the tip or blade of the tongue close to the alveolar ridge. The air escapes along the centre of the tongue (Roach, 2009). The teeth function as an obstacle which produces additional turbulence. Therefore, the noise created is relatively intense and generates a loud hissing sound, compared to the labiodental and dental fricatives.

/ʃ/ and /ʒ/ are postalveolar fricatives, meaning the blade of the tongue is close to the postalveolar ridge area (Roach, 2009). Similar to /s/ and /z/, the teeth act as an obstacle,

[1] That the vocal tract can act as a filter is based on the source-filter theory of speech production. For a detailed explanation, see Stevens (2000).

which produces extra turbulence noise. However, unlike for /s/ and /z/, the blade of the tongue moving behind the alveolar ridge creates a small room between the teeth and the bottom of the tongue. This room is called sublingual cavity. The sublingual cavity "adds length to the front cavity of the vocal tract, and thus lowers its resonance frequency" (Johnson, 2011, p. 159). Therefore, although the place of articulation of /s, z/ and that of /ʃ, ʒ/ are adjacent to each other, these two groups of fricatives sound quite different. In fact, these two places of articulation mark the distinction [+anterior] and [−anterior] (Evers et al., 1998; F. Li, Edwards, et al., 2007: F. Li, Munson, et al., 2011).

In addition, lip rounding (or lip protrusion) has an acoustic effect of lowering the formant frequencies due to the enlarged or extended oral cavity (Johnson, 2011, p. 159). Likewise, when pronouncing an alveolar fricative followed by the close back rounded vowel /u/, the lips are also rounded as a result of coarticulation. Therefore, the alveolar fricative in this phonetic environment sounds similar to a postalveolar fricative.

3.2 The acoustic properties of fricatives

As mentioned briefly in Section 3.1, the overall spectrum of a fricative is determined by the size and shape of the front cavity, and whether there is an obstacle to produce additional turbulence noise. Although the eight English fricatives are distinct in terms of place of articulation and voicing, there is not an acoustic cue which can single out a fricative from the others (Jongman et al., 2000). Nevertheless, previous studies have identified a number of spectral, amplitudinal, and temporal properties of the frication noise which can distinguish different English fricatives (Jongman et al., 2000; Maniwa et al., 2009; Nissen and Fox, 2005). Spectral properties include peak, slope, centre of gravity, standard deviation/variance, skewness, kurtosis, and F2 onset frequency. Amplitudinal properties include normalised root-mean-square amplitude, relative amplitude, and harmonics-to-noise ratio (or called harmonicity). Temporal properties include normalised noise duration. In the following sections, each acoustic property of fricatives is discussed with respect to the spectral, amplitudinal, and temporal aspect. The effects of place of articulation, gender, voicing, and vowel from previous studies are also reported.

3.2.1 Spectral properties

Peak location

The peak is defined as the highest location of a wave. Previous studies showed that sibilants have distinct peaks while there are no well-defined peaks in the spectra of non-sibilants (Maniwa et al., 2009; Jongman et al., 2000). The spectra of non-sibilant fricatives are also relatively flat, compared to that of sibilant fricatives. Jongman et al. (2000) measured the spectral peak as the highest amplitude of a FFT spectrum (with a pre-emphasis factor of 98%). All the acoustic analyses in their study were conducted on the production of 20 American English speakers. All the recordings were sampled at 22 kHz with a 16-bit quantization and 11 kHz low-pass filter applied. The audio signals were transformed using the fast Fourier transform (FFT) algorithm and a 40 ms Hamming window was generally applied to extract the acoustic features. The acoustic properties were then subjected to a four-way analysis of variance (ANOVA) (i.e. place x voicing x vowel x gender), which was basically a factorial ANOVA. It was found that labiodental fricatives had the highest peak value (7733 Hz), followed by dental fricatives (7470 Hz), then by alveolar fricatives (6839 Hz) and postalveolar fricatives (3820 Hz). Bonferroni *post hoc* tests showed that all four places of articulation were significantly different from each other (Jongman et al., 2000).

Jongman et al. (2000) also found a main effect of voicing in that voiceless fricatives had a significantly higher spectral peak value than voiced fricatives, and an interaction effect of place and voicing. *Post hoc* tests suggested that "the difference in spectral peak location between voiceless and voiced fricatives was carried by the non-sibilant fricatives" but not the sibilant fricatives (Jongman et al., 2000, p. 1256). There was also a main effect of gender that the spectral peak value of female speakers was significantly higher than that of male speakers. As for the following vowel, there was no main effect but there was an interaction effect of place and vowel. *Post hoc* tests revealed that the spectral peak for /s, z/ was significantly lower if the following phonetic environment was back rounded vowels (/o, u/).

In the study by Maniwa et al. (2009), the fricative production of 20 American English speakers were examined. The goal of their study was to compare the fricatives in clear speaking and conversational styles. The speech signal was converted using discrete Fourier transform (DFT), which was presumably FFT, and a 20 ms Hamming window. Regarding the spectral peak, it was defined as "the frequency bin corresponding to the largest value in [the spectrum] $X(f)$" (Maniwa et al., 2009). The mean peak frequency was measured for the centre three window locations in the DFT spectra with a pre-emphasis factor of 98%. Mixed-model factorial ANOVAs (style x fricative x gender) were then computed. They found that generally labiodental and dental fricatives had a high mean peak frequency below 10 kHz (Maniwa et al., 2009). The mean peak frequency of postalveolar fricatives was particularly low (around 3500-4000 Hz) in comparison with alveolar fricatives.

Both Jongman et al. (2000) and Maniwa et al. (2009) suggested that the mean spectral peak frequency decreases as the place of articulation moves further back in the oral cavity. Moreover, the postalveolar fricatives generally have a lower/mid-frequency spectral peak than the anterior fricatives. This is due to the sublingual cavity and the enlarged front cavity when producing postalveolar fricatives (Johnson, 2011). It also explains why the peak value of the alveolar fricatives /s, z/ is lower when followed by back rounded vowels than the front vowels. In summary, the spectral peak location is effective in distinguishing all four places of articulation. Voicing, gender, and place x vowel are potential predictors.

Centre of gravity

The centre of gravity, which is also called the first spectral moment or M1, is the average energy concentration of a spectrum. Jongman et al. (2000) calculated the spectral moments at four different window locations (onset, middle, end, offset) using a 40 ms Hamming window. The offset window was the last 20 ms of the fricative and the first 20 ms of the next sound. They found that the centre of gravity of FFT spectra was highest for alveolar fricatives (6133 Hz), followed by labiodental (5108 Hz) and dental fricatives (5137 Hz), and was lowest for postalveolar fricatives (around 4229 Hz). The difference of centre of gravity was not significant enough to differentiate between labiodental and dental fricatives. That is to say, the mean values of centre of gravity of sibilants and non-sibilants were significantly different, and the differences between alveolar and postalveolar fricatives were also significant. Nevertheless, the centre of gravity at the second and fourth window locations were able to distinguish all four places of articulation (Jongman et al., 2000).

Nissen and Fox (2005) examined the acoustic characteristics of voiceless (General American) English fricatives /f, θ, s, ʃ/ by young children. Only the monosyllabic words with an initial syllable CV(C)(C) were investigated. Similarly, the recordings were sampled at 44.1 kHz using a 16-bit quantization and a 22.05 kHz low-pass filter. The spectral moments were measured only for those fricatives longer than 80 ms. Three 40 ms windows located at the beginning, middle, and end of the fricative was applied to extract the measurements. Each window was pre-emphasised by first-differencing. The signals were converted using FFT algorithm with a fixed number of points (n=2048). Zero-padding, which is a technique used to increase the length of FFT by adding zeros, was applied when necessary. The centre of gravity values were transformed to the Equivalent Rectangular Bandwidth (ERB) scale before running the statistical analysis. ERB is a psychophysical metric, "which employs a 'notched-noise' method rather than traditional masking procedures to measure the auditory filter bandwidth of the human auditory system" (Nissen and Fox, 2005, p. 2572). ANOVA results indicated that the centre of gravity differed significantly across place of articulation. *Post hoc* results showed that the significant difference held for all comparisons except between /θ/ and /ʃ/. It is surprising to see such a finding since one would expect that the difference was not significant between non-sibilants (/f, θ/) but not between sibilant fricative /ʃ/ and non-sibilant fricative /f/. It could be the case that there is a mistype of symbol in the original article. In the general discussion, spectral mean was said to be able to separate non-sibilant from sibilant fricatives, and also between sibilant fricatives (/s, ʃ/).

It is generally expected that there is a correlation between the length of the front cavity and the resonating frequency. F. Li, Munson, et al. (2011, p. 1001) explained that "the longer the front resonating cavity is, the lower the overall resonating frequencies in the fricative spectrum will be, which is reflected in a lower M1 value". This prediction is partially confirmed except between the labiodental and dental fricatives (Jongman et al., 2000; Maniwa et al., 2009; Nissen and Fox, 2005). Theoretically speaking, the vowel context may also have an effect on the centre of gravity of the fricatives due to coarticulation effects. For example, Nittrouer et al. (1989) showed that the centre of gravity of the voiceless alveolar fricative in close front unrounded vowel /i/ contexts was higher than in close back rounded vowel /u/ contexts. However, the following vowel was not a significant indicator of centre of gravity in the study by Jongman et al. (2000).

Regarding other factors, both Jongman et al. (2000) and Maniwa et al. (2009) reported that voiceless fricatives generally had a higher center of gravity than voiced fricatives. This effect was particularly evident in non-sibilants (Maniwa et al., 2009). A main effect of gender was also found that female speakers had a higher centre of gravity than male speakers (Jongman et al., 2000).

Standard deviation/Variance

The standard deviation of a spectrum, which is also called the second spectral moment or M2, refers to how dispersed the spectrum is from the mean frequency. In the study by Jongman et al. (2000), the variance was reported instead of the standard deviation, which is the square root of the variance. Place of articulation had a main effect. *Post hoc* results indicated that the variance for non-sibilants (6.37 MHz for /f, v/ and 6.19 MHz for /θ, ð/) was significantly larger than for sibilants (2.92 MHz for /s, z/ and 3.38 MHz for /ʃ, ʒ/). Among the sibilants, the postalveolar fricatives had a significantly larger variance than the alveolar fricatives. There was no significant difference between the variance of labiodental fricatives and dental fricative. Nevertheless, when taking the window locations into account, the variance distinguished all places of articulation except for the second window location (Jongman et al., 2000).

Similar to Jongman et al. (2000), Nissen and Fox (2005) also found a main effect of place of articulation. *Post hoc* results showed that all four places of articulation were significantly different with respect to variance. The descending order of variance values was: labiodental fricative (6.26 MHz), dental fricative (5.38 MHz), alveolar postalveolar fricative (3.30 MHz), and fricative (2.39 MHz). Unlike Jongman et al. (2000), the variance difference between non-sibilant fricatives was significant in the study by Nissen and Fox (2005).

In addition, a main effect was found for voicing that voiced fricatives had a greater variance than voiceless fricatives but the effect size was relatively small ($\eta^2 = 0.069$) (Jongman et al., 2000). Female speakers were also found to have a larger variance than male speakers (Jongman et al., 2000).

Skewness

Skewness is a measure of the asymmetry of the spectral shape (i.e. the spectral tilt). It is also called the third spectral moment or M3. Zero skewness suggests a symmetrical distribution. White noise is an example which has zero skewness (Boersma and Weenink, 2004). Positive skewness indicates "a negative tilt with a concentration of energy in the lower frequencies", whereas negative skewness suggests "a positive tilt and a predominance of energy in the higher frequencies" (Jongman et al., 2000, p. 1253). Jongman et al. (2000) found that skewness distinguished all four places of articulation in fricatives. It was difficult to comprehend the mixed results of skewness values in the study by Jongman et al. (2000), namely why labiodental fricatives (0.077) and postalveolar fricatives (0.693) had positive skewness (i.e. left tilted), whereas dental fricatives (-0.083) and alveolar fricatives (-0.229) had negative skewness (i.e. right tilted). It would be expected that among non-sibilant fricatives and among sibilant fricatives, there would be a similar concentration of energy. In terms of absolute value, the ascending order according to place of articulation was: labiodental, dental, alveolar, and postalveolar. It can only be concluded that the distribution of energy was mainly concentrated on the lower frequencies for postalveolar fricatives.

Nissen and Fox (2005) also found a main effect of place of articulation for skewness. The effect came from the postalveolar fricative. Unlike the findings in the study by Jongman et al. (2000), the skewness values for /f/ (-2.23), /θ/ (-2.18), /s/ (-1.88) were all negative, meaning they were positively tilted, except for /ʃ/ (0.21), meaning it was negatively tilted to the lower frequencies.

Voicing and gender were reported to have a main effect on skewness. It was found that voiceless fricatives generally had a higher skewness value than voiced fricatives; and male speakers generally had a larger skewness than female speakers (Jongman et al., 2000; Maniwa et al., 2009).

Kurtosis

The spectral kurtosis, which is also called the fourth spectral moment or M4, is a measure of tailedness. It denotes "how much the shape of the spectrum around the centre of gravity is different from a Gaussian shape" (Boersma and Weenink, 2004, p. 780). The kurtosis of a standard normal distribution is three. Therefore, it is conventional to subtract three when reporting the kurtosis (such as in Maniwa et al., 2009). Following this measurement, positive kurtosis suggests heavy tails; whereas negative kurtosis indicates light tails.

Jongman et al. (2000) found a main effect of the place of articulation. Labiodental (2.11) and alveolar (2.36) fricatives had the highest kurtosis, followed by dental fricatives (1.27) and postalveolar fricatives (0.42). All pairwise comparisons were significant, except between labiodental and alveolar fricatives. Nissen and Fox (2005) reported slightly different results. It was found that although there was a main effect of place of articulation, the effect mainly came from the particularly small kurtosis value of postalveolar fricatives (1.46), while the values were similar for labiodental (3.78), dental (3.77) and alveolar (3.54) fricatives.

In terms of voicing, Jongman et al. (2000) found that voiceless fricatives had larger kurtosis than the voiced fricatives but the effect size was small ($\eta^2 = 0.001$). There was also an effect of gender that female speakers had a higher value of kurtosis than the male speakers (Jongman et al., 2000).

F2 onset

Jongman et al. (2000) measured the second formant (F2) at the following vowel onset using a 23.3 ms Hamming window. ANOVA results revealed a four-way interaction effect of place of articulation, voicing, vowel, and gender. Nevertheless, it was complicated to interpret what such an interaction actually implied. Jongman et al. (2000) did not discuss this interaction in detail. There was a main effect of place of articulation on the estimation of F2 onset frequency. A general pattern was that F2 onset values increased as place of articulation moved further back in the vocal tract. Nevertheless, the difference between dental and alveolar fricatives was not significant. There was a place x vowel interaction and the F2 onset values differed significantly for labiodental and alveolar fricatives. Another significant interaction was between voicing and place of articulation: the F2 onset value was significantly higher for dental and postalveolar fricatives when followed by /i, e/. In addition, there was a main effect of vowel in that the mean F2 values were higher for front vowels (/i, e, æ/) than back vowels (/u, o, ɑ/), and the F2 values increased when the vowel height increased. All the comparisons among vowels were significant except between /o/ and /ɑ/. A main effect of gender was found and *post hoc* tests showed that the F2 onset value was significantly higher for females than males.

Slope

Jongman et al. (2000), Maniwa et al. (2009), and Nissen and Fox (2005) all measured slope but of different parts in the spectrum. Jongman et al. (2000) was more interested in the locus equation, which is a measurement of the transition information based on the F2 onset and midpoint value of the vowel. The slope and y-intercept were calculated, following the method in the study by Sussman, McCaffrey, et al. (1991) and Sussman, Hoemeke, et al. (1993). It was found that only the slope of labiodental fricatives was significantly different from other fricatives; while the y-intercept of labiodental and postalveolar fricatives were distinctive, the y-intercept of dental and alveolar fricatives was not (Jongman et al., 2000). An effect of gender was found with the y-intercept for females being higher than for males.

Maniwa et al. (2009) measured the two spectral slopes, namely one low-frequency slope and one high frequency slope, following the method by Jesus and Shadle (2002). The average peak frequency was first computed in the logged DFT spectrum with a pre-emphasis factor of 98%. Two linear regression lines were fitted based on the average peak frequency (i.e. below the peak frequency and above the peak frequency till 15 kHz) using the least square method. It was found that the low-frequency slope was higher for sibilant fricatives than non-sibilant fricatives, and the low-frequency slope of postalveolar fricatives was higher than that of alveolar fricatives.

Nissen and Fox (2005) measured the slope by fitting a linear regression line based on a fixed range of frequency (i.e. 1-15 kHz) in the power spectrum. There was an effect of place of articulation for the spectral slope. The slope values in ascending order by place of articulation were: 3.40 for /θ/, 3.42 for /f/, 5.46 for /s/, and 9.08 for /ʃ/. *Post hoc* tests revealed that the spectral slopes between sibilant fricatives and non-sibilant fricatives, and between sibilant fricatives were significantly different from each other. The difference between non-sibilant fricatives was not significant. The small values of non-sibilant fricatives confirm previous findings that the spectra of non-sibilant fricatives are relatively flat. Also, the slope of postalveolar fricatives was significantly steeper than alveolar fricatives.

3.2.2 Amplitudinal properties

Normalised root-mean-square amplitude

Root-mean-square amplitude is the amplitude of a sound signal multiplied by the square root of the mean square. Jongman et al. (2000) measured the normalised root-mean-square amplitude of fricatives by subtracting the vowel root-mean-square amplitude from the root-mean-square amplitude of the entire fricative. The vowel amplitude was defined as the root-mean-square amplitude "averaged over three consecutive pitch periods at the point of maximum vowel amplitude" (Jongman et al., 2000, p. 1256), following the same approach by Behrens and Blumstein (1988). By doing so, the intensity differences among speakers were normalised. A main effect of place of articulation was found for the normalised amplitude. *Post hoc* tests showed that the amplitudes of all four places of articulation were significantly different from each other (Jongman et al., 2000). In terms of magnitude, postalveolar fricatives were the loudest, followed by alveolar fricatives, and then by labiodental fricatives and dental fricatives. This result is in line with the mechanical model of fricative production as described in Section 3.1. There was also an effect of voicing that voiceless fricatives had a significantly larger amplitude. An effect of vowel context was also reported but the effect mainly came from the difference between /i/ and /ɑ/. It indicated that the even after normalising the root-mean-square amplitude by using the vowel root-mean-square amplitude, the effect of vowel was still significant. In addition, there was an interaction effect of place and voicing and *post hoc* tests revealed that the amplitude difference between voiced and voiceless non-sibilant fricatives was significantly higher than for voiced and voiceless sibilant fricatives. There was no significant effect of gender.

Similarly, Nissen and Fox (2005) measured the normalised amplitude also by subtracting the root-mean-square amplitude of the entire frication segment from that "of the strongest component within the initial 40 ms of the following vowel" (p. 2572). An effect of place of articulation was found and *post hoc* results indicated that the effect came from the differences between non-sibilant and sibilant fricatives. Unlike the findings in the study by Jongman et al. (2000), there was not much difference between /f/ (-13.7 dB) and /θ/ (-11.9 dB), and between /s/ (-3.6 dB) and /ʃ/ (-3.0 dB). There was also an effect for vowel in that the normalised amplitude values significantly decreased when the following vowel was /ɑ/. An interaction effect of place and vowel was found and *post hoc* results indicated that the following vowel /ɑ/ and /f, θ, s/ had an interaction effect but not /ɑ/ and /ʃ/.

Relative amplitude

Jongman et al. (2000) and Maniwa et al. (2009) measured the relative amplitude by subtracting the vowel amplitude from the fricative amplitude. Unlike the root-mean-square amplitude, the vowel amplitude was measured as the third formant (F3) region for sibilants and as the fifth formant (F5) region for non-sibilants, while the fricative amplitude was measured at the centre of the fricative. The findings in the study by Jongman et al. (2000) were complicated. In general, there was a main effect of place of articulation and *post hoc* tests demonstrated that the relative amplitudes of all four places were different from each other. Alveolar fricatives had the largest amplitude, followed by dental fricatives, then by labiodental fricatives, and lastly by postalveolar fricatives. There were also interactions between place and vowel, place and voicing, and place and gender (Jongman et al., 2000).

Harmonics-to-noise ratio

Harmonics-to-noise ratio measures the ratio between periodic (harmonics) and aperiodic (noise) components of a speech sound (Fernandes et al., 2018). In the study by Maniwa et al. (2009), the harmonics-to-noise ratio was calculated by measuring the difference between the amplitude of the periodic part of the fricative and the amplitude of the noise of the fricative. Maniwa et al. (2009) found that the harmonics-to-noise ratios of voiced fricatives were higher than those of the voiceless fricatives. Moreover, the harmonics-to-noise ratios of voiced non-sibilants were higher than those of the voiced sibilants.

3.2.3 Temporal properties

Normalised noise duration

Noise duration is the length of the frication noise. Instead of the absolute duration of the frication noise, Nissen and Fox (2005) and Jongman et al. (2000) calculated the normalised noise duration using the ratio of fricative duration over word duration. They argued that the absolute duration may vary, based on the speaking rate. Jongman et al. (2000) found a main effect of place of articulation. The mean duration by place of articulation in ascending order was: 0.333 ms for labiodental fricatives, 0.340 for dental fricatives, 0.382 for alveolar fricatives, and 0.393 for postalveolar fricatives. *Post hoc* tests showed that all duration differences were significant except between labiodental and dental fricatives. There was also a main effect of voicing in that voiceless fricatives were significantly longer than voiced fricatives. An interaction effect between place and voicing was found: the effect of voicing was more evident for non-sibilant fricatives than sibilant fricatives. Regarding gender, there was a main effect that male speakers had a longer normalised duration than female speakers.

Nissen and Fox (2005) measured the absolute duration of the fricative segments. It was found that place of articulation had a main effect and it was due to the decreased duration of /f/. Interestingly, there was an effect of vowel and *post hoc* tests suggested that the duration decreases significantly when followed by /ɑ/ comparing that by /i/.

In summary, most spectral properties were significantly different i) between sibilants and non-sibilants, ii) between alveolar fricatives and postalveolar fricatives, but not between dental fricatives and labiodental fricatives (except for spectral peak and normalised root-mean-square amplitude). Gender and voicing were predictor variables for most of the spectral properties, while vowel was also an indicator for F2 onset frequency and normalised root-mean-square amplitude. Several interactions among place of articulation, voicing, gender, and vowel were reported. The amplitudinal and temporal properties were able to distinguish the place of articulation and voicing. Harmonics-to-noise ratio was particularly useful to distinguish between voiced and voiceless fricatives of the same place of articulation. The normalised root-mean-square amplitude, the relative amplitude, and noise duration were only able to distinguish sibilants from non-sibilants. Vowel was reported an significant indicator of the amplitude but it was only limited to certain vowel contexts. All these findings were based on (standard) American English and it would be interesting to see if these acoustic properties were able to distinguish Hong Kong English fricatives with respect to place of articulation and voicing.

3.2.4 DCT coefficients

As mentioned briefly in Section 3.2.1, to measure the aforementioned acoustic and spectral properties, the majority of the studies convert the speech signals into a DFT (amplitude) spectrum using the FFT algorithm, which is a fast version of Fourier Transform. Fourier transformation is a decomposition process of a finite signal sequence from the time-domain. The output is a representation in the frequency domain composed of a set of complex sinusoids at integer cycles (i.e. $k = 0, 1, 2, ..., N - 1$) (Harrington, 2010). Theoretically speaking, the summation of the sinusoids can reconstruct the original signal. There is another strand of study on the acoustic and spectral properties using the Discrete Cosine Transformation (DCT) (amplitude) spectrum. The DCT analysis of fricatives has been gaining more and more popularity over the past two decades (see Bukmaier and Harrington, 2016; Guzik and Harrington, 2007; Harrington, 2010; Jannedy and Weirich, 2017; Stuart-Smith, 2020). The DCT algorithms decompose a signal into a set of cosine waves at half cycles (i.e. $k = 0, 0, 5, 1.0, 1.5, ..., 1/2(N - 1)$) (Harrington, 2010), unlike the full integer cycles in FFT. Moreover, the output of the DCT is a set of sinusoids with no phase, which equals to a cosine wave. Therefore, this operation is named discrete *cosine* transformation. For these studies (Bukmaier and Harrington, 2016; Guzik and Harrington, 2007; Jannedy and Weirich, 2017; Stuart-Smith, 2020), DCT coefficients were measured and discussed with respect to the features of fricatives. DCT coefficients are the resulting amplitudes of the cosine waves at the respective frequencies (Jannedy and Weirich, 2017). The zeroth coefficient (k_0) is the amplitude of the cosine wave at frequency $k = 0$, and is proportional (but not equal) to the signal's mean; the first coefficient (k_1) is the amplitude of the cosine wave at frequency $k = 0.5$, and is inversely proportional to the signal's slope; the second coefficient (k_2) is the amplitude of the cosine wave at frequency $k = 1$, and is proportional to the signal's curvature (Harrington, 2010; Jannedy and Weirich, 2017). The list continues till the $(N - 1)$th coefficient, which is the frequency $k = 0.5(N - 1)$. Most studies only reported and discussed the first three or four coefficients as they encode more global and general properties of the signal shape. In the following, the methods and findings from previous studies on the acoustic aspect of fricatives, though not limited to English fricatives, are reviewed.

Jannedy and Weirich (2017) studied the acoustic differences between the voiceless palatal fricative /ç/ and the voiceless postalveolar fricative /ʃ/ of different German speaker groups using by examining both the four spectral moments (i.e. centre of gravity, standard deviation, skewness, and kurtosis) and the first four DCT coefficients (i.e. k_0-k_1). 130 speakers from three different groups were recruited to read aloud a word list with minimal pairs embedded in a carrier phrase. The recordings were sampled at 48 kHz. The spectral analysis was limited to 500 Hz to 12 kHz. The DCT coefficients were extracted after the spectral frequency were converted to the Bark scale. Linear mixed effect models were computed with the Euclidean distances between the fricatives in the four-dimensional spectral moments space as the dependent variable and repetition, speaker group, minimal pair, and their interactions as the fixed effects. A random intercept for the speaker and a by-speaker random slope for minimal pair were applied. A forced choice perception test from another group of 12 participants was also conducted. It was found that the DCT analysis mirrored the perception results better than the spectral moment analysis and was able to capture the slight differences observed in the spectral shapes. It demonstrated that DCT coefficients were a more reliable measurement of the acoustic properties than the spectral moments using window functions. Jannedy and Weirich (2017, p. 404) explained that it was because "the DCT coefficients quantify the entire shape of the spectrum rather than just the central frequency or the weighting of the higher or lower frequencies".

Stuart-Smith (2020) investigated the production of the voiceless alveolar fricative /s/ and voiceless postalveolar fricative /ʃ/ in the Glasgow dialect and the gender effect over time. Word-initial target fricatives from interviews and causal conversations were examined using the DCT analysis. Some of the reasons for adopting DCT analysis are stated as follows (Stuart-Smith, 2020, p. 3):

- DCT analysis can be applied to tracks of variable length, without arbitrary decisions about how many points to take for a track, or time normalisation (Watson and Harrington, 1999).
- DCT coefficients are continuous, mathematically-independent measures, amenable to linear mixed modelling to test for the influence of fixed and random factors, separately and in interaction.
- agent-based modelling of sound change which reduces acoustic trajectories, also for /s/-retraction, to a three-point multidimensional space using DCT analysis is proving effective (Harrington and Schiel, 2017; Harrington, Kleber, et al., 2018).

It is interesting to see how the static and dynamic measurements were performed by Stuart-Smith (2020). First, the static centre of gravity was calculated using a 10 ms Hamming window, which was applied in the central 70% of the fricative segment, in a presumably FFT spectrum. It was very different from the 40 ms full Hamming window adopted in the study by Jongman et al. (2000). The static spectral slope was computed in a Long-Term Average Spectrum (LTAS), which "represents the logarithmic power spectral density as a function of frequency" (Boersma and Weenink, 2004, p. 631). The dynamic measurements were calculated by first extracting the track of values of centre of gravity using a sequence of 10 ms Hamming window in the central 70% of the fricative segment. Then, the DCT coefficients were calculated using the tracks of centre of gravity measures. The same process was applied to slope. This approach was similar to Watson and Harrington (1999), in which the formant trajectories or contours of vowels in Australian English were compressed into DCT coefficients. In the study by Watson and Harrington (1999), it made sense to probe into the DCT coefficients of the each extracted formant trajectory in a vowel segment. Nevertheless, it was hard to understand what the DCT coefficients of a sequence of extracted spectral means and spectral slopes represented in the study by Stuart-Smith (2020). Since the DCT coefficients are already a form of compression with dynamic information of the spectrum encoded, methodologically, it would be more straightforward to compute the DCT coefficients from the whole FFT fricative segment and to interpret the first coefficient corresponding to the sequence's mean and the second coefficient corresponding to the sequence's slope, as demonstrated in (Jannedy and Weirich, 2017). Similar to the study by Jannedy and Weirich (2017), all the static and dynamic measures of centre of gravity and slope were estimated using the mixed linear regression models with respect to gender and age group in the study by Stuart-Smith (2020).

Bukmaier and Harrington (2016) studied the physiological and acoustic characteristics of three sibilant fricatives /s, ʂ, ɕ/ in Polish from nine speakers. A Mel-scaled DCT-transformation was adopted. First, the speech signals were converted to a 256 point DFT spectrum "with a 40 Hz frequency resolution, 5 ms Blackman window, and a frame shift of 5 ms" (Bukmaier and Harrington, 2016, p. 5). Then, the DFT spectrum's frequency axis was converted into Mel. After that, the DCT coefficients were derived from the Mel-scaled DFT spectrum at the midpoint of the frication. Harrington (2010) explained that converting the Hz scale into an auditory scale (e.g. Bark or Mel) was not just because the scale was "more closely related to the way in which the frequency [was] perceived" by human ears, but also because the Mel-scaled coefficients such as the Mel Frequency Cepstral Coefficients (MFCC) were used more frequently in automatic speech recognition; mel-scaled coefficients were needed "to distinguish effectively between different phonetic categories than when DCT coefficients [were] derived from a Hz scale" (Harrington, 2010, p. 213).

In the present study, both acoustic properties of the FFT spectrum and the DCT coefficients (k_0, k_1, k_2, and k_3) of fricatives are examined. Most previous studies adopted the DCT analysis of sibilant fricatives, and it would be interesting to see if this analysis can be extended to study non-sibilant fricatives. Since the DCT analysis is dependent on the spectral shape, and previous studies using other acoustic characteristics already demonstrated that it is hard to distinguish between the spectra of labiodental and dental fricatives, it is questionable if DCT analysis is suitable for studying non-sibilant fricatives. Nevertheless, the findings in the study by Jannedy and Weirich (2017) showed that the slight difference in the spectral shape could be captured by the DCT coefficients but not the spectral moments. The DCT analysis might be able to show differences between non-sibilant fricatives as well.

This study examines the following acoustic and spectral properties: the four spectral moments (centre of gravity, standard variation, skewness, kurtosis), spectral peak, spectral slope, normalised amplitude, F2 onset frequency, harmonic-to-noise ratio, as well as the first four DCT coefficients (k_0, k_1, k_2, k_3). The advantages of using the Mel-scale mentioned by Harrington (2010) for automatic speech recognition can be best manifested when using them together with a log power spectrum, which is actually MFCC. Therefore, it does not make a huge difference whether the DCT coefficients at this stage is Hz-scaled or Mel-scaled. For the acoustic analysis, this study sticks with the Hz-scaled DCT coefficients. The algorithms and methods adopted for each acoustic property are discussed in Section 6.3.3. The effects of place of articulation, voicing, and gender are also examined.

3.3 Classification of fricatives

One of the goals of conducting acoustic analysis is to find out the relevant acoustic features and build a classification model that can be comparable to human perception. Therefore, many studies conducted a classification analysis of fricatives using the acoustic characteristics being investigated (e.g. Forrest et al., 1988; Jongman et al., 2000; Nissen and Fox, 2005; Bukmaier and Harrington, 2016).

One common classification method noted in previous studies is discriminant analysis, which is similar to a regression analysis except that the output variable is a discrete and mutually exclusive class, usually by the four places of articulation (i.e. quadratic discriminant analysis). A set of linear equations or discriminant functions are derived in discriminant analysis (Nissen and Fox, 2005). The discriminant algorithms model the conditional distribution of y given x $p(x|y)$. Jongman et al. (2000) conducted a step-wise linear discriminant analysis using 21 predictors (i.e. spectral peak location, the four spectral moments x the four window locations, F2 onset frequency, normalised root-mean-square amplitude, relative amplitude, and normalised duration). The jackknife resampling technique was applied, "whereby each speaker in turn was used as the testing speaker with training being done on the 19 remaining speakers" (Jongman et al., 2000, p. 1260). The final classification scores were averaged across the 20 testing speakers. Preliminary results showed that the accuracy rates were 88% for sibilant fricatives and 66% for non-sibilant fricatives. The standardised canonical discriminant function coefficients were analysed to evaluate the predictive effect of each input variable. A new classification model was built using the suggested features. And these two procedures continued until a best model was built. The final classification accuracy was 53% for labiodental fricatives, 48% for dental fricatives, 81% for alveolar fricatives, and 88% for postalveolar fricatives. It would also be interesting to know the classification rate in terms of voicing since Jongman et al. (2000) reported a main effect of voicing for many acoustic properties. Unfortunately, Jongman et al. (2000) did not include it in their paper.

Nissen and Fox (2005) adopted a similar approach described in the study by Jongman et al. (2000). The predictor variables were frication noise duration, normalised root-mean-square amplitude, spectral slope and the four spectral moments (i.e. centre of gravity, standard deviation, skewness, and kurtosis). Validation results in the training model indicated that the classification rates for sibilant fricatives were 95%, and 70% for non-sibilant fricatives. Error analysis revealed that confusions mainly came from misrecognising /θ/ as /f/ or /f/ as /θ/ but "rarely crossed the sibilant/non-sibilant distinction" (Nissen and Fox, 2005, p. 2577).

Other studies adopted a Gaussian classification, which is a generative classifier. A generative learning algorithm models the $p(x|y)$ and $p(y)$ as opposed to $p(y|x)$, and often Bayes Rule (see Equation 3.1) is used to predict the conditional probability distribution $p(y|x)$. A class posterior is calculated based on the class conditional density (which is assumed to have a Gaussian distribution) and the class prior.

$$p(y|x) = \frac{p(x|y)p(y)}{p(x)} \tag{3.1}$$

Bukmaier and Harrington (2016) built a Gaussian model to classify the Polish fricatives of nine speakers using the two Mel-scaled DCT coefficients (k_1 and k_2). Leave-one-out cross-validation was applied, and the process was similar to the jackknife method in the studies by Jongman et al. (2000) and Nissen and Fox (2005). The difference was that in the leave-one-out cross-validation, "a given speaker's data were classified following training on the data of the other eight speakers" (Bukmaier and Harrington, 2016, p. 6) and is iterated for all speakers in turn, while the jackknife method computed statistics only from the kept samples. Results indicated that in the slow speech rate, the classification accuracy was 96% for /s/, 77% for /s̺/, and 63% for /ɕ/. There was a high degree of confusion for /ɕ/ and /s̺/.

Abdelatty Ali et al. (2001) studied the effect of acoustic-phonetic features for the automatic classification of fricatives. First, the input fricative was classified into voiced or voiceless based on the duration threshold of the unvoiced portion of the frication. Then the fricative was further classified into three places of articulation, namely alveolar, postalveolar, and dental (combining both labiodental and interdental) using five main features: i) the maximum normalised spectral slope, ii) the location of the most dominant spectral slope,

iii) the location of the most dominant spectral peak, iv) the spectral centre of gravity, and v) the dominance relative to the highest filter, which "describes the amplitude roll-off in the high-frequency filters relative to the dominant peak" (Abdelatty Ali et al., 2001, p. 2228). The model was clear and intelligible as knowledge-based decision-tree-like algorithms were employed. Statistical discriminant analysis and Bayesian classification, which used the maximum posterior probability (similar to the approach described in the study by Bukmaier and Harrington (2016)) was carried out. The overall classification accuracy for voicing was 93%. The overall classification accuracy for place of articulation was 91%. The overall classification accuracy of fricatives was 87%.

To further evaluate the acoustic-phonetic feature-based model, an artificial neural networks (ANN) using 12 MFCCs as input features was built for comparison in the study by Abdelatty Ali et al. (2001). MFCCs were obtained by applying the DCT to the filter banks and only keeping some of the resulting coefficients while the rest being discarded. Multilayer perceptron, which is a class of feedforward artificial neural network, and the back-propagation algorithm were used to build the network. There was one input layer, one hidden layer, and one output layer. Results showed that the the accuracy rate for voicing classification was 77% and that for place of articulation classification was 86%. The findings suggested that use of acoustic-phonetic feature-based approach could improvement classification performance. One possible explanation was that the number of input features were more for the ANN, while some of the features might not be relevant for the classification of voicing and place of articulation of fricatives, and hence, created confusion for the algorithms. On the other hand, the input acoustic features were carefully selected when modelling.

In recent years, there are more uses of convolutional neural networks (CNNs) and recurrent neural network (RNN) (e.g. Anjos et al., 2020; Arora et al., 2018; Patgiri et al., 2013). Convolution neural networks take the name from a mathematical linear operation between matrices called convolution (Albawi et al., 2017). A CNN uses convolution to extract high level-features, usually used in image recognition but can also be applied to speech and voice recognition.

A basic CNN architecture is sketched in Figure 3.3 (Phung and Rhee, 2019). Basically, the convolutional layer extracts the features from the input data using the mathematical operation of convolution. A filter (also called kernel) is applied and shifts through the input data to perform this convolutional operation. The output of the convolutional layer is a learnt feature map. A pooling layer decreases the size of the convolved feature map. Usually, max pooling is performed which takes the largest element from the feature map. The pooling layer helps reduce the computational costs. A fully connected layer primarily performs the classification. It also carries the weights and biases of the features and the classes. A weight decides how important the input is between two neurons. All of a neuron's inputs are multiplied by their weights and added together before being turned into an activation value. A bias value, which is a constant, is an additional input into the next layer. It helps ensure there is always an activation in the neuron regardless of the input values. In the classification stage, the number of neurons in the final output layer is the same as the number of classes or labels. For model training, two hyper-parameters, namely epoch and batch size, need to be defined. An epoch defines how many times the learning algorithm will work through the entire training dataset, and a batch size decides the number of samples to be propagated through the network.

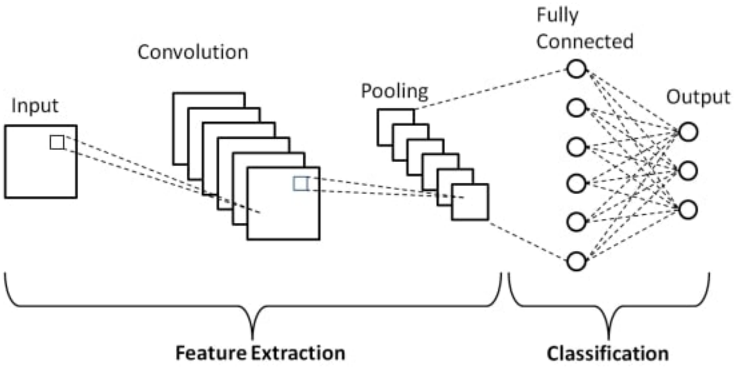

Figure 3.3 Schematic diagram of a basic convolutional neural network (CNN) architecture (Phung and Rhee, 2019, p. 3)

In the study by Anjos et al. (2020), two models using CNN and log Mel filter banks were built to classify fricatives by place of articulation (labiodental, alveolar, and postalveolar) and by voicing (voiced and voiceless). The models were trained on European Portuguese speech samples from 356 children.The same architecture was deployed for both networks. The input features were matrices of log Mel filter banks, from which MFCCs could be computed, and they were 80 (bins) x 9 (frames) matrices extracted by a 25 ms window function and 10 ms shift size. There were two convolutional layers, each followed by a pooling layer. The first convolution layer used 50 kernel filters (size: 10 x 2) and a 2 x 1 stride. Max pooling with a 2 x 2 window (stride = 1) was then applied. The second convolutional layer used 25 kernel filters (stride = 1), followed by max pooling. The output pooled feature maps were flattened before feeding into the fully connected network with four hidden layers (with 1000, 500, 100, and 10 neurons respectively). The output layers have three neurons for place of articulation and two neurons for voicing. "The models were trained for 100 epochs, with a batch size of 10" (Anjos et al., 2020, p. 3158). The overall accuracy for place of articulation was 90.4% and that for voicing was 90.9%. The overall F1-score for place of articulation was 88.0% for labiodental fricatives, 87.6% for alveolar fricatives, and 93.1% for postalveolar fricatives. The overall F1-score was 83.3% for voiced fricatives and 93.8% for voiceless fricatives.

To conclude, different classification methods and algorithms were reviewed in this section. Regardless of which method, it was generally found that the classification accuracy of non-sibilant fricatives was lower than that of sibilant fricatives. Also, the accuracy rate was lower for voiced fricatives than voiceless fricatives. Regarding the input features, the main difference between different approaches is whether the features are imposed by the researcher or learnt. Deep neural networks, in general, are more suitable to learn more complex representations in the data. Nevertheless, it is widely known that one of the disadvantages of deep neural networks is that the internal logic is opaque. It is unlike the decision-tree-like model demonstrated in the study by Abdelatty Ali et al. (2001). For the purpose of the present study, classification models using convolutional neural network were built and applied for the classification of fricatives and their variants.

Chapter 4

Automatic speech recognition (ASR) and the Munich AUtomatic Segmentation (MAUS) system

4.1 General review of feature-based ASR systems

Automatic speech recognition (ASR) systems differ a lot, especially in recent decades in which End-to-End (E2E) systems are gaining more popularity. Nevertheless, many mainstream ASR systems are still using the statistical acoustic-language model approach. The mechanism of this approach is briefly introduced in this section in order to have an overall understanding of the ASR pipeline and what the function of each component is in this process.

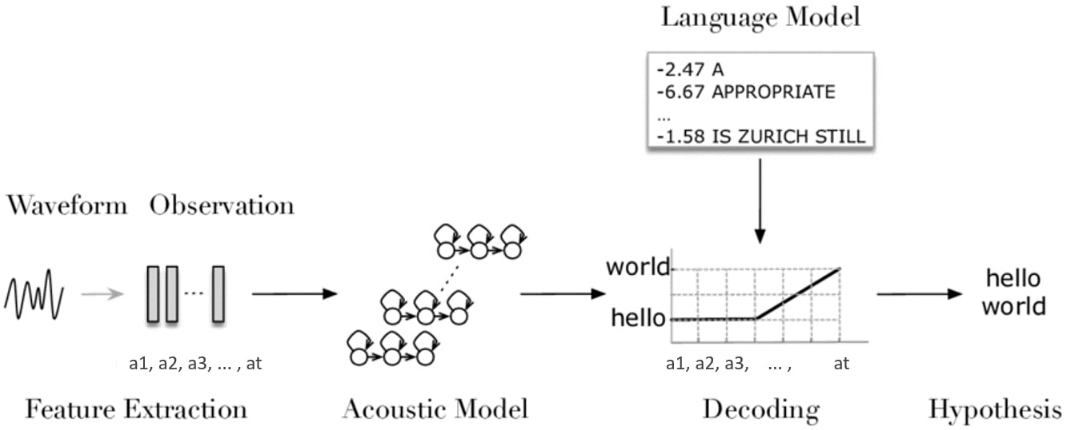

Figure 4.1 A statistical ASR system adapted from C. Zhang (2017, p. 10)

The statistical approach of ASR generally involves an acoustic model, a language model, and a decoder, as demonstrated in Figure 4.1. First, the speech signal is converted into a sequence of acoustic vectors, $A = a1, a2, a3, ..., a_t$. Each vector contains representations of the speech signal of a certain period of time (a_t). One common type of acoustic features is the Mel-scale frequency cepstral coefficients (MFCCs), as already mentioned in Section

3.2.4. How MFCCs are computed is described in Section 6.6.1. An alternative to MFCCs is perceptual linear prediction (PLP). The language model proposes the sequence of n words $W = w1, w2, w3, ..., w_n$ and the sequence gets converted into basic speech units such as phones based on the pronunciation dictionary which provides the prior probability of each word sequence. According to Jelinek (1997), if $P(W|A)$ denotes the probability of the words W being spoken, given the acoustic evidence A observed, the function of ASR can be presented as follows:

$$\hat{W} = \arg\max_{w} P(W|A) \tag{4.1}$$

Basically, the *argmax* operation returns the word sequence of the highest probability, which is also considered as the "maximum *a posteriori* (MAP) decoding rule" (C. Zhang, 2017, p. 10). The Bayes' Rule (see also Section 3.2.4) can be applied so that:

$$P(W|A) = \frac{P(W)P(A|W)}{P(A)} \tag{4.2}$$

Since the maximisation in Equation 4.1 is using a fixed acoustic sequence A, *P(A)* in this case is a constant. Therefore, Equation 4.1 and Equation 4.2 can be simplified as:

$$\hat{W} = \arg\max_{w} P(W)P(A|W) \tag{4.3}$$

The *a prior* probability of observing a word sequence $P(W)$ is determined by the language model which is usually an n-gram model. A pronunciation dictionary, which contains the pronunciations of words and tagged attributes is often used in the process to convert words into speech units. The probability of observing the acoustic sequence A given the word sequence W (i.e. $P(A|W)$ is determined by the acoustic model which usually consists of hidden Markov models (HMMs) (Jelinek, 1997). The scores of the language model and acoustic model are further processed in the decoder, which consists of finite state transducers (FSTs), "to estimate the final output in the form of sequences of phonemes, words, or sentences" (Arora et al., 2018, p. 99).

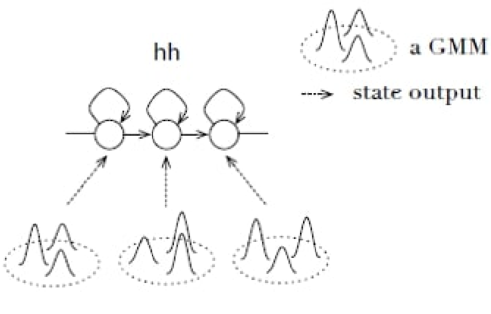

Figure 4.2 A schematic representation of GMM-HMM for the phone hh, cited from C. Zhang (2017, p. 18).

Specifically, in a phone-based acoustic model, the probabilistic distribution of the features for a phone can be modelled with a Gaussian Mixture Model (GMM), whereas the transition between phones and the respective acoustic features of the speech signal can be modelled with the Hidden Markov Model (HMM) (Hui, 2019). It is assumed that "the properties of its associated signals can be characterised by a discrete time Markov process or Markov chain" (C. Zhang, 2017), which is also known as discrete-time Markov chain (DTMC). It means that the speech signal can randomly change its state in time and each state is one of the finite distinct states. The change of state is based on the probability distribution stored in the current and preceding states. This kind of model is called GMM-HMM and is often seen in the literature. A schematic representation of a GMM-HMM is illustrated in Figure 4.2.

Another crucial component of a feature-based ASR system is the language model. As mentioned earlier, the language model is usually an n-gram model which stores the *prior* probability that $P(W)$ of each word sequence W dependent on $n-1$ preceding word (C. Zhang, 2017). The product $P(W)$ can be formally presented as Equation 4.4 (Jelinek, 1997, p. 57), where $P(w_i|w_1, ..., w_i)$ is the probability w$_i$ will be spoken given the words spoken before $w_1, ..., w_i$.

$$P(W) = \prod_{i=1}^{n} P(w_i|w_1, ..., w_i) \tag{4.4}$$

The number of n differs in each language model and different n yields different ASR results. It is typical to use a 3-gram or 4-gram model. What needs to be noted is that each n-gram unit is devised based on the training data, which usually comprises multiple corpora. Unseen n-gram units with zero frequency in the training data can cause errors in the ASR output $\hat{W}$ regardless of the acoustic signal, due to Equation 4.3 (C. Zhang, 2017).

The acoustic model, pronunciation dictionary, and language model are integrated in the decoder, which creates a finite state graph structure. In technical terms, the task of ASR is to find the most likely path in the created graph that "will have generated the utterance to be decoded according to the MAP decoding rule" in Equation 4.3 (C. Zhang, 2017, p. 24). To do so, the *Viterbi* algorithm (Viterbi, 1967) is often employed. The *Viterbi* path is the maximum *a posteriori* probability estimate of the maximising state sequence of the hidden states (Jelinek, 1997, p. 23).

To sum up, this section provides an overview of a feature-based ASR system, which uses a statistical approach. Major components, namely the acoustic model and the language model, were reviewed. The common algorithms and techniques were also briefly introduced. Although the above discussions are primarily based on word recognition, the theories form the backbone of many state-of-the-art forced alignment tools (see Section 4.2).

4.2 Forced alignment and MAUS

Forced alignment refers to the process of which orthographic transcriptions are automatically aligned to the time intervals in the audio files to generate the sentence, word, or phone level segmentation. It is different from speech recognition which generates the text transcription based on the audio file. In phonetics research, often the phone segmentation and phone recognition are of interest. For example, since the present study is interested in analysing the relation between fricatives and the corresponding signal (e.g. acoustic and spectral signal) as well as the relation between the realisation of fricatives and linguistic factors, phone segmentation and phone labelling are crucial for this study. Nevertheless, manually segmenting the phone and auditorily labelling each phone is very time-consuming, especially when the dataset is large. A forced alignment tool is required.

There are many open-source forced alignment tools, as shown in Table 4.1 (Pettarin, 2018). As can be seen, most forced alignment tools use the hidden Markov model (HMM) as the base statistical model and the HMMs are created using the hidden Markov toolkit (HTK) which supports density mixture Gaussians (GMMs). It can be assumed that most tools are based on speech recognition algorithms, such as those mentioned in Section 4.1.

Table 4.1 Open-source forced alignment tools (Pettarin, 2018)

Name	Algorithm	Supported Language(s)	Interface
aeneas	DTW	30+	CLI, LIB, Web
CMU Sphinx	HMM (own), RNN	11	CLI, LIB
DARLA	HMM (HTK)	English	Web
FAVE-align	HMM (HTK)	English	CLI, (Web)
Gentle	HMM (Kaldi)	English	CLI, Web
Julius	HMM (own)	English, Japanese	CLI, LIB
Kaldi	HMM (own), DNN, RNN	English	CLI, LIB
kaldi-dnn-ali-gop	HMM(Kaldi), DNN(Kaldi nnet3)	English	CLI, LIB
LaBB-CAT	HMM (HTK)	English	Web
MAUS	HMM (HTK)	21	CLI, Web
Montreal Forced Aligner	HMM (Kaldi)	English	CLI
Penn Forced Aligner (P2FA)	HMM (HTK)	English	CLI, Web
Prosodylab-Aligner	HMM (HTK)	English	CLI
SailAlign	HMM (HTK)	English, Greek, Spanish	CLI
SPPAS	HMM (Julius)	12+	CLI, GUI

CLI: command line interface
DTW: Dynamic Time Warping
DNN: Deep Neural Network
GUI: graphical interface
HMM: Hidden Markov Model
HTK: Hidden Markov Toolkit
LIB: library callable by third party software
RNN: Recurrent Neural Network

Among the state-of-the-art forced alignment tools, the Munich Automatic Segmentation System (MAUS) (Schiel, 1999) was adopted for the present study because it allows a high degree of custom configuration. The custom configuration is suitable for processing different types of speech data as well as testing different linguistic hypotheses as demonstrated by Kisler et al. (2017). Nevertheless, the command line interface (CLI) of MAUS, which is in C language, only supports German. The English models are only available in the web interface. Therefore, WebMAUS (Kisler et al., 2017), which is the web service for MAUS, was employed.

The workflow of MAUS is illustrated in Figure 4.3. As can be seen, the processes in

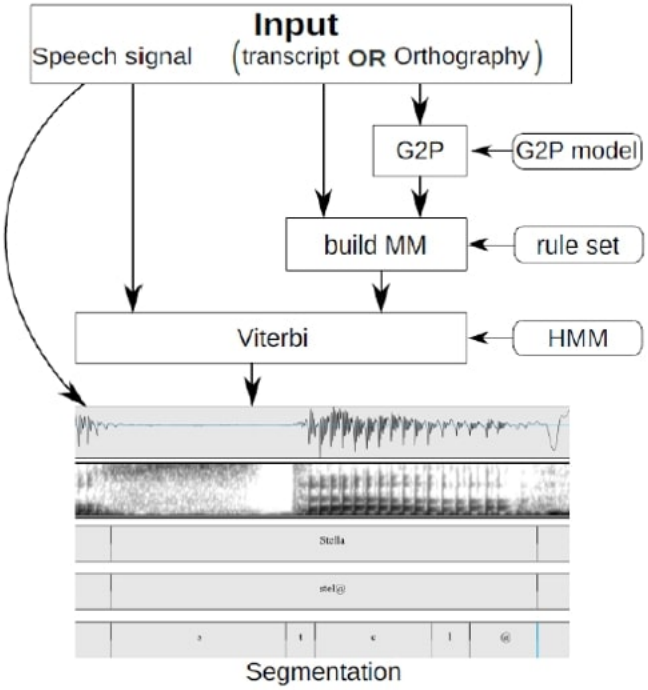

Figure 4.3 Workflow of MAUS cited from Kisler et al. (2017)

MAUS are similar to the ASR mechanism mentioned in Section 4.1. If the text is an orthographic transcript, it passes to the Grapheme-to-Phoneme (G2P) model which transforms the text into the "most-likely standard pronunciation combined with a number of (optional) annotations" (Kisler et al., 2017, p. 14). It mainly relies on pronunciation dictionary lookup. Classifiers which were trained using decision tree algorithms are also applied. If the input text is a phonetic transcription in Speech Assessment Methods Phonetic Alphabets (SAMPAs), the G2P process is not necessary. The phones from the G2P model or SAMPA transcription are then passed to a language specific Markov model (MM). The MM calculates the probabilities of all pronunciation variants for a given canonical phone symbol. This can be done by applying the "statistically weighted rewrite rules" based on the language specific rule set (Kisler et al., 2017, p. 11). The rewrite rules are automatically learnt from a large (phonetically transcribed) corpus. How the rewrite rules are generated is discussed in greater detail in Section 4.3.

The pronunciation variants and the conditional probabilities are then "transformed into a Markov process, in which the nodes represent phonetic segments and the arc between them represent transition probabilities" (Kisler et al., 2017, p. 11). An example of an *a priori* pronunciation Markov Model of MAUS for the German word ‘Abend’ (‘evening’) is illustrated in Figure 4.4. As can be seen, if there is only one transition path, the probability is 1 (e.g. /ʔ/→/aː/, /n/→/t/, /m/→/t/). If there are more than one paths, the sum of all transition paths is always 1 (e.g. /aː/→/b/+/aː/→/m/ and /b/→/ə/+/b/→/n/+/b/→/m/). The Markov model is then passed to the decoder.

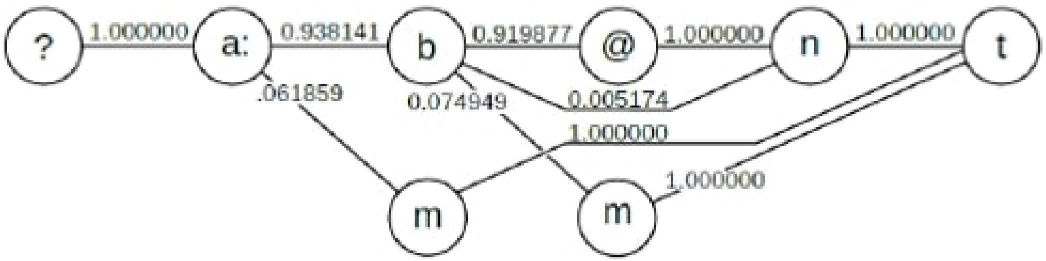

Figure 4.4 *A priori* pronunciation Markov Model of MAUS for the German word 'Abend' ('evening') cited from Kisler et al. (2017, p. 38). The phone symbols are SAMPAs.

The decoder is actually a hidden Markov Model (HMM) in which the underlying Markov chain is usually hidden. MAUS uses the *Viterbi* alignment, which is the maximum likelihood alignment, as the estimate. The pre-processed speech signals are also passed to the decoder. The MFCCs features ($n = 12$) are first extracted from the input audio files. Apart from the MFCCs, MAUS also uses Energy and the first and second derivative (presumably of the spectrum) (Kisler et al., 2017) as the acoustic features. The decoder uses the backtracking algorithm to search for the most probable path. This is how the phone segmentation and labelling are conducted in MAUS.

Almost all components of the ASR system can be configured in MAUS except the SAMPA phone symbols which have to conform to the phonemic symbols recognised by MAUS for a particular language. For example, as shown in Figure 4.3, MAUS supports two types of input transcription: the orthographic transcription and the SAMPA phonemic transcription. A rule set file and the phone insertion probability (INSPROB) can be added which directly affects the probability of the transition path in the Markov chain. The pron model weight, which gives more weights to the scores of the pronunciation model than the acoustic model, can be specified. Higher weight means the path of canonical pronunciation from the pronunciation dictionary is more likely to be selected, whereas lower weight means that the path of pronunciation from the acoustic model (based on the acoustic evidence) is more likely to be selected.

In the present study, the SAMPA phonemic transcription was used as the input transcription for the acoustic analysis and the pronunciation rule set for Hong Kong English fricatives was devised. For the purposes of comparison, the same pron model weight was used. Formulating the probabilistic pronunciation rule set is not a trivial task. In the next Section, the MAUS rule set is discussed in more detail using the example of the British English rule set.

4.3 The MAUS pronunciation rule set

In MAUS, the purpose of a pronunciation rule is to constrain the graph in the Markov model (e.g. Figure 4.4 in Section 4.2). Therefore, the probability of each pronunciation rule needs to be specified. MAUS uses two types of pronunciation rule set (Schiel, 2015): i) without statistical information (hand-crafted) and ii) with statistical information (data-driven). For the former type, the probability is based on the findings from the literature and empirical studies. If previous studies did not provide any statistical information, it is hard to calculate the probability. Therefore, the rule probabilities of this type are, by default, set to 1. For the latter type, the *a posteriori* probability is derived from the observation frequency in the corpus data (typically 1 hour of speech) (Schiel, 1999). Whether MAUS uses the former or latter rule set is dependent on whether there are sufficient annotated training data for that language. In the language inventory in WebMAUS, details of the rule set is generally listed for each language.

The general syntax of a pronunciation rule in the rule set with statistical information is as follows (Schiel, 1999):

$$L, B, R > L, N, R; P \tag{4.5}$$

where *L, B, R* and *L, N, R* are sequences of SAMPA symbols, and *L* and *R* denote the left and right context of *B* and *N* respectively. The left and right context must be exactly one symbol. *P* is the rule probability: $P(L, B, R|L, N, R)$. Each symbol is separated by a comma. In this rule, *B* is replaced by *N*. To denote the word boundary, usually in the case of word-initial or word-final phone, the hashtag '#' can be used. To denote the beginning of an utterance, '<' can be used. If the SAMPA symbol contains a backslash, the backslash has to be rewritten as '-' (ASCII 45) (Schiel, 2021). If the SAMPA symbol contains a digit (e.g. '2:' and '9:' in the German inventory), it has to be preceded by 'P'. The P, here, is a real alphabet, whereas the *P* in Equation 4.5 denotes probability. The rule probability *P* uses natural log.

Here is an example of a pronunciation rule in MAUS (Schiel, 2021):

$$d, n =, \# > d, @, n, \# \quad -0.693100 \quad 0.000000 \tag{4.6}$$

This rule denotes that if the word-final syllabic nasal /n=/ is preceded by /d/, it will be replaced by a schwa /@/ (IPA: /ə/) followed by the alveolar nasal /n/. The *a prior* probability of this rule is -0.6931. Assuming there is only one rule given the condition (*d,n=,#*), the probability of the *P(match|match)* rule is 1 - *p(replacement|match).* This rule does not have to be set explicitly as it is automatically computed in MAUS.

In the rule set without statistical information, the general syntax of a pronunciation rule is as follows:

$$..., L - B - R, ... > ..., L - N - R, ... \tag{4.7}$$

The *a prior* probability does not need to be written explicitly since it is assumed to be 1 for each rule set. It also implies that there can only be one matching and replacement condition per rule. The rule set without statistical information uses '-' as the context separator. Hence, any SAMPA symbol which contains a backslash (which is encoded as '-' as well) cannot be used. Nevertheless, this rule set has less constraints on the number of SAMPA symbols. For example, more SAMPA symbols can be added on top of the left and right context, as shown in Equation 4.8. The left and/or right context can also be not specified at all, as shown in Equation 4.9 (Schiel, 2021) .

$$aI - C - s, t, \# > aI - k - s, t, \# \tag{4.8}$$

The above rule replaces /C/ (IPA: /ç/) with /k/ in /aICst/ (IPA: aɪçst) in word final position.

$$- N, k- > -N, g- \tag{4.9}$$

The above rule replaces /Nk/ (IPA: /ŋk/) with /Ng/ (IPA: /ŋg/) at any arbitrary position.

Both types of rule set have their own advantages and disadvantages. If there are enough data, the statistically weighted rule set is generally preferred as it yields better performance for the pronunciation model (Schiel, 2015). Therefore, the present study attempts to create statistically weighted rules for Hong Kong English fricatives based on the auditory analysis results and the phone recognition results are evaluated.

Chapter 5

Research questions

The present study seeks to answer the following research questions (RQs) with respect to the phonetics and phonology of Hong Kong English fricatives:

RQ1 a) Which fricatives can be found in Hong Kong English and what are their distributions in terms of frequency?

RQ1 b) Which variants of fricatives can be found in Hong Kong English and what are their distributions in terms of frequency?

RQ2 a) Which acoustic properties of Hong Kong English fricatives can distinguish all four places of articulation (i.e. labiodental, dental, alveolar, and postalveolar) and voicing (i.e. voiced and voiceless)? Are these acoustic properties for classification the same as those for Inner Circle English fricatives?

RQ2 b) What are the acoustic characteristics of Hong Kong English fricatives? Do they share the same pattern as the Inner Circle English fricatives?

RQ3 Which linguistic factors (i.e. syllable position, stress, preceding labial consonants, preceding /u/, and following /u/) influence the realisation of Hong Kong English fricatives?

RQ4 To what extent can the findings of this study be applied to an existing state-of-the-art automatic speech recognition (ASR) system and improve the phone recognition of Hong Kong English fricatives and their variants?

Regarding RQ1a, it is predicted that the voiceless fricatives /f, θ, s, ʃ/ will be mainly found in Hong Kong English, as suggested by Hung (2000). While the realisations of /f, s, ʃ/ will be primarily the same as the phonemic representation, it is predicted that only two-third of the realisations of /θ/ will be [θ]. The remaining one-third will be variants of /θ/ (Deterding et al., 2008; Hansen Edwards, 2019; Hung, 2000). As for the voiced fricatives /v, ð, z, ʒ/, it is predicted that their occurrences will be marginal.

Regarding RQ1b, it is predicted that the main variants of /v/ will be the voiceless [f] and the voiced labialised velar approximant [w] (Bolton and Kwok, 1990; Hung, 2000). As for /θ/, the variants will include [f] and [s], as reported by Hansen Edwards (2019). Specifically, the majority of the variants of /θ/ will be [f] and only a small proportion of variants will be [s]. /z/ and /ʒ/ will be unanimously realised as the voiceless counterparts [s] and [ʃ] respectively. Although Bolton and Kwok (1990) observed that /ʃ/ was also realised as [s], other studies did not seem to share the same view (Hung, 2000; Setter et al., 2010). In the current study, it is predicted that not much variation will be observed for /ʃ/, meaning /ʃ/ will be mostly realised as [ʃ].

Regarding RQ2a, it is predicted that similar to previous studies on (standard) American English fricatives (Jongman et al., 2000; Maniwa et al., 2009; Nissen and Fox, 2005), most acoustic properties will be able to distinguish between sibilant and non-sibilant fricatives, but not between non-sibilant fricatives (/f, v, θ, ð/). Only a small number of properties will be able to distinguish all four places of articulation. Since the acoustic properties which distinguished all four places were different in previous studies on (standard) American English fricatives, for the present study, it is predicted that the acoustic properties will also be different from those reported in previous studies. With regards to voicing, previous studies found that most acoustic properties were able to distinguish between voiced and voiceless fricatives (Jongman et al., 2000; Maniwa et al., 2009; Nissen and Fox, 2005). Therefore, it is predicted that voicing is also a main predictor for most acoustic properties of Hong Kong English fricatives.

As for RQ2b, no previous studies have suggested that there is an acoustic difference between the fricatives in Hong Kong English and Inner Circle Englishes. Hence, it is predicted that the patterns of acoustic properties with respect to place of articulation will be the same. For example, the spectral peak value will be highest for labiodental fricatives, followed by dental fricatives, then by alveolar fricatives, and lastly by postalveolar fricatives in Hong Kong English.

With respect to RQ3, it is hypothesised that syllable position will be a main predictor for the realisation of Hong Kong English fricatives. TH-fronting is more likely to occur in syllable onset position than in coda position, as suggested by Hansen Edwards (2019). It is also hypothesised that TH-fronting is more likely to occur when there is a preceding labial consonant in the same word. The realisation of /s/ and /z/ will be more susceptible to the co-articulation effect from the following vowel. It is speculated that /s/ and /z/ are more likely to be pronounced as [ʃ] when followed by back rounded vowels (Johnson, 2011). The realisation of /ð/ as [d] (TH-stopping) is more likely to occur in syllable onset position and the variant [f] is more likely to occur in syllable coda position, as proposed by Bolton and Kwok (1990). As for stress, it is speculated that stress will have no effect on the realisation of fricatives based on the findings reported by Hansen Edwards (2019).

With respect to RQ4, since it is expected that systematic variation of the realisation of Hong Kong English fricatives can be found, the findings of RQ3 can be transformed into phonological rules with *a priori* probabilities and applied to an actual ASR system. As there is not yet a well-trained acoustic model of Hong Kong English and based on the predictions in RQ2b, an existing model of English (e.g. standard British English) can be adapted. It is predicted that with the adaptation, the phone recognition of Hong Kong English fricatives and their variants will be improved.

Chapter 6

Method

6.1 Participants

106 students (F51M55, mean age 20.5) from tertiary institutions in Hong Kong were recruited as the participants for this study. All of them were raised and educated in Hong Kong and spoke Cantonese as their first and dominant language. They also learnt English and Mandarin in primary and secondary school. They did not have any long-term study abroad experience in English-speaking countries. Table 6.1 summarises the information of the participants[1].

Table 6.1 Summary of the participants' background information

N = 106	*n*	%
Gender		
-Female	51	48.1
-Male	55	51.9
Birthplace		
-Hong Kong	102	96.2
-Mainland China	4	3.8
English proficiency		
-High (HKDSE 5*/5**)	19	17.9
-Mid (HKDSE 4/5)	72	67.9
-Low (HKDSE 2/3)	15	14.2

The participants of this study can be interpreted as a group of educated speakers of English in Hong Kong or "mid-range" Hong Kong English speakers (Bolton and Kwok, 1990; Hung, 2000; Q. Zhang, 2013, see also Section 2.4), who were raised and educated in post-colonial Hong Kong. The composition of the participants also makes this study comparable with previous studies (e.g. Bolton and Kwok, 1990; Jim YH Chan, 2013; Deterding et al., 2008; Hansen Edwards, 2019; Hung, 2000), which also examined the English production of university students in Hong Kong.

[1] The categorization of English proficiency was based on the Hong Kong Diploma of Secondary Education Examination (HKDSE) English scores of the participants

6.1.1 Subset of participants

The 106 participants were randomly shuffled by gender and each participant was assigned a speaker ID (from F01 to F51 for females, and from M01 to M55 for males). The data of 25% of the participants (no. of speakers = 27) were used for acoustic analysis and training the classification models. The subset of participants was gender-balanced (14F13M). Table 6.2 shows the information of each participant in the subset. All of them were born in Hong Kong. The mean age was 21.4. The participants came from different faculties: 33.3% of them studied science, 18.5% of them studied arts and humanities, 14.8% of the participants studied social science, engineering, and business respectively, and 3.8% studied Medicine. In terms of English proficiency, 74.1% were mid, 18.5% were high, and 7.4% were low.

Table 6.2 Information of the participants in the subset

Speaker ID	Age	Birthplace	Faculty	English Proficiency
F01	22	Hong Kong	Social science	Mid
F02	21	Hong Kong	Medicine	Mid
F03	21	Hong Kong	Science	High
F04	22	Hong Kong	Business	High
F05	19	Hong Kong	Engineering	Mid
F06	20	Hong Kong	Business	High
F07	22	Hong Kong	Business	Mid
F08	20	Hong Kong	Arts and Humanities	Mid
F09	22	Hong Kong	Business	Mid
F10	19	Hong Kong	Science	High
F11	22	Hong Kong	Science	Mid
F12	20	Hong Kong	Arts and Humanities	Mid
F13	22	Hong Kong	Arts and Humanities	Mid
F14	26	Hong Kong	Social science	Mid
M01	21	Hong Kong	Engineering	Low
M02	22	Hong Kong	Science	High
M03	22	Hong Kong	Social science	Mid
M04	21	Hong Kong	Science	Mid
M05	20	Hong Kong	Arts and Humanities	Mid
M06	19	Hong Kong	Science	Mid
M07	20	Hong Kong	Science	Mid
M08	28	Hong Kong	Engineering	Mid
M09	24	Hong Kong	Science	Mid
M10	20	Hong Kong	Science	Mid
M11	19	Hong Kong	Social science	Mid
M12	19	Hong Kong	Engineering	Low
M13	24	Hong Kong	Arts and Humanities	Mid

6.2 Procedures and materials

6.2.1 Data collection procedures

The data collection took place in February and March 2019 at two universities in Hong Kong. To observe the regulations of conducting research studies which involve human subjects in Hong Kong, ethics approval granted by Human Research Ethics Committee was obtained prior to the recruitment of participants and data collection. The participants were recruited through permitted advertisements using mass mailing and posters on campus. All participants were informed that participation was completely voluntary and they could withdraw from the study anytime. Informed consent form was signed before the start of any task. Each participant was rewarded $100 Hong Kong Dollar as reimbursement upon completion of all tasks.

The participants were recorded individually at a sound-proof booth inside a language laboratory. Condenser microphone Audio-Technica AT2020USB+ was used. The microphone was placed in front of a monitor and around 15-20 cm away from the speaker. It was pivoted approximately 45 degrees to the mouth of the speaker to prevent turbulence from direct airflow (Jongman et al., 2000). The recording programme Audacity (Version 2.2.2) was adopted. The sample rate (or sampling rate) was set to 48 kHz and the bit depth was set to 16-bit. The higher sample rate resulted in more measurements per second, and hence, better reconstruction of the original audio. Given that this study aimed to measure fricatives, which are high frequency noises, setting the sample rate to 48 kHz allowed the Nyquist frequency up to 24 kHz, which provided further buffer before downsampling and filtering (see Section 6.3.3).

The participants were instructed to read aloud a word list and a story. Details of the reading materials are illustrated in Section 6.2.2. The materials were embedded in a PowerPoint and presented on the monitor. For the word list, each PowerPoint slide contained one sentence. As for the story, each slide contained one paragraph. Once the participants finished a slide, they clicked the pointer to proceed to the next slide at their own pace. There was a practice for the word list using non-target words so that the participants were able to familiarise themselves with the format and procedures. After the practice, the researcher left the sound-proof booth, and the participants began the tasks. The data collection was around 60 minutes per participant and there were three 5-minute breaks in between. The whole process was invigilated by the researcher via the control system connected to the sound-proof booth in the language laboratory. The participants could communicate with the researcher at any time using the intercom.

6.2.2 Materials

Word list

To collect the production of target fricatives systematically, a list of pseudo-words was generated. A 2 × 4 factorial design was adopted with two syllable positions (onset, coda) and four canonical vowels (/ɪ, e, u, a/). Each syllable had a structure of an onset consonant, a vowel, and a coda consonant: C_1VC_2. The consonant was either a target fricative (/f, v, s, z, θ, ð, ʃ, ʒ/), a plosive (/p, b, t, d, k, g/), or the labiovelar approximant /w/ in each syllable. The CVC syllables were combined to form a disyllabic pseudo-word with the stress always on the first syllable. To prevent by-design voicing and devoicing assimilation (Abdelli-Beruh, 2012), the adjacent consonants (i.e. the coda of the first syllable and the onset of the second syllable) had the same voicing pattern.

Each pseudo-word was embedded in a carrier phrase: *Say* _____ *again* and each sentence was read twice by the participants. In the word list presented to the participants (see Appendix A), prototypical spelling representation of the phonemes in English were employed except for the fricatives /θ, ð, ʒ/. To avoid confusion with the spelling of /t, d, z/, the corresponding IPA symbols were used. The participants were given time to learn the pronunciation of the three symbols on their own, based on how they pronounced some real English words. The same applied to the four vowels. There were four sets of word list, each had a different randomised order of sentences.

256 target fricatives (32 tokens per fricative) were produced by each participant. In total, 27,136 tokens of fricatives were collected from the 106 participants.

Story

The reading passage (see Appendix B) was based on the story *Snow White and the Seven Dwarfs*, originally by Brothers Grimm. The story was chosen because it is a well-known fairy tale both worldwide and in Hong Kong. The participants were familiar with the story line. The source passage (Ashliman, 1996) was shortened and modified by the researcher by including more target fricatives. In the PowerPoint slides presented to the participants, there was a picture beside each paragraph to facilitate comprehension. Altogether, there were 653 words and 2267 phones in the story. 448 phones (19.8%) were fricatives. The distribution of each fricative is listed in Table 6.3. The frequency counts were based on the (standard) British English phonemic transcription of the words in the story. Compared to the word list, the phonetic environments of the fricatives were more diverse in the story. For example, there were more different vowels (monophthongs and diphthongs). Various consonant clusters were also included. Unlike the word list, the frequencies of the fricatives in the story were not equally distributed.

What is of more interest in the story data is the +/- 1 preceding and following context of the target fricative. Altogether, there were 194 unique phoneme sequences. 58 of them were in word-initial position, 89 were in word-medial position, and 47 were in word-final position. Table 6.4 displays the sequences per fricative in SAMPA symbols of MAUS, separated by commas (see Section 4.3 for the encoding). The story data were used to evaluate and validate the results from the word list.

Table 6.3 No. of fricatives in the story

Fricative	*n*	Words (unique and sorted per fricative)
f	48	after, beautiful, beef, coffin, dwarfs, fairy, famous, fell, felt, fetch, fever, find, first, fish, five, fled, following, food, footpath, for, forest, forks, frame, fresh, frightened, from, front, funeral, further, herself, knife, life, paragraphs, refined, refused, safety, therefore
v	56	advice, advised, alive, arrived, attractive, develop, devise, ever, everyone's, evil, fever, five, gave, give, given, have, having, however, invited, live, lived, love, moved, movie, native, of, saved, seven, seventeenth, survived, vegetables, venture, version, very, view, village, vinegar, voice, voodoo
θ	32	authors, birth, birthday, death, earth, footpath, health, months, mouth, pythons, seventeenth, south, sympathy, thanks, thought, thousand, threatened, three, through, throwing, truth, with, worth
ð	74	breathed, brothers, clothing, further, neither, that, the, them, then, there, therefore, they, this
s	130	advice, also, answer, asked, assault, assumed, crisis, crossed, decided, devise, disguised, distant, dressing, dwarfs, east, famous, first, forest, forks, glass, herself, huntsman, inside, instantly, its, juicy, just, kiss, lets, lips, mixed, months, most, once, outside, paragraphs, passage, passed, past, poisonous, prestige, prince, princess, response, sabotaged, safety, said, same, saved, saw, screaming, screwed, season, send, seven, seventeenth, shrimps, situation, skin, small, snow, so, soaked, someone, soon, soup, south, special, spell, spoons, spread, spring, stay, still, story, strangers, strawberry, suicide, surprise, survived, sympathy, task, thanks, this, voice, whites
z	69	advised, always, anyways, as, authors, because, brothers, cheese, crazy, days, design, disguised, Disney, easily, everyone's, example, is, noise, owners, pictures, poisonous, presumably, pythons, realised, refused, season, spoons, strangers, surprise, thousand, using, vegetables, was, whose, wizard, zigzag, zipped, zombies, zoned
ʃ	31	condition, ensure, fish, fresh, mentioned, mushroom, rubbish, rushed, share, she, shocked, shoot, shouted, shrimps, situation, special, sure, wish, wished
ʒ	8	Asia, genre, leisure, prestige, rouge, sabotaged, usual, version
total	448	

Table 6.4 Unique +/- 1 phoneme sequences found in the story

Fricative	Sequence (displayed in SAMPA symbols of MAUS)
f	#,f,3: #,f,I #,f,O: #,f,Q #,f,U #,f,aI #,f,e #,f,e@ #,f,eI #,f,i: #,f,j #,f,l #,f,r #,f,u: A:,f,s A:,f,t I,f,@ I,f,aI I,f,j O:,f,s Q,f,I aI,f,# e@,f,O: eI,f,t f,s,# i:,f,# l,f,#
v	#,v,3: #,v,I #,v,OI #,v,e #,v,j #,v,u: @,v,aI I,v,# I,v,@ I,v,aI I,v,d I,v,e Q,v,# V,v,# aI,v,# aI,v,d d,v,aI e,v,@ e,v,r eI,v,# eI,v,d i:,v,# i:,v,@ n,v,aI u:,v,I u:,v,d {,v,# {,v,I
θ	#,T,O: #,T,aU #,T,r #,T,{ 3:,T,# 3:,T,d @,T,I A:,T,# I,T,# O:,T,@ T,s,# aI,T,@ aU,T,# e,T,# l,T,# n,T,# n,T,s u:,T,#
ð	#,D,@ #,D,I #,D,e #,D,e@ #,D,eI #,D,{ 3:,D,@ @U,D,I V,D,@ i:,D,@ i:,D,d
s	#,s,@ #,s,@U #,s,I #,s,O: #,s,V #,s,aU #,s,e #,s,eI #,s,i: #,s,k #,s,m #,s,n #,s,p #,s,t #,s,u: #,s,{ 3:,s,e 3:,s,t @,s,# @,s,Q @,s,j @U,s,t A:,f,s A:,s,# A:,s,k A:,s,t I,s,# I,s,aI I,s,g I,s,p I,s,t O:,f,s OI,s,# Q,s,t T,s,# V,s,t aI,s,# aI,s,I e,s,# e,s,I e,s,t f,s,# i:,s,t k,s,# k,s,t l,s,@U n,T,s n,s,# n,s,@ n,s,aI n,s,e n,s,t p,s,# t,s,# t,s,aI t,s,m u:,s,I {,s,I
z	#,z,@U #,z,I #,z,Q #,z,u: @,z,# I,z,# I,z,@ I,z,aI I,z,j I,z,n I,z,u: OI,z,# OI,z,@ Q,z,# aI,z,# aI,z,d aU,z,@ e@,z,# eI,z,# eI,z,I g,z,A: g,z,{ i:,z,# i:,z,@ i:,z,I l,z,# m,z,# n,z,# u:,z,# u:,z,I u:,z,d {,z,#
ʃ	#,S,Q #,S,U@ #,S,aU #,S,e@ #,S,i: #,S,r #,S,u: I,S,# I,S,@ I,S,t V,S,r V,S,t e,S,# e,S,@ eI,S,@ n,S,@ n,S,O:
ʒ	#,Z,Q 3:,Z,@ A:,Z,d e,Z,@ eI,Z,@ i:,Z,# u:,Z,# u:,Z,U

, denotes the phone delimiter

# denotes the word boundary

6.3 Data processing pipeline for acoustic analysis

The collected word list data were first processed using a pipeline. The idea of establishing a pipeline is to automate the steps for conducting acoustic analysis. The pipeline in this study is limited to preparing the speech data for acoustic analysis and the training data for classification. Statistical analysis is not included in the pipeline. A full automation of the pipeline cannot be achieved due to the necessity of manual correction of the phone boundaries and phonetic labels. Figure 6.3 is a graphical outline of the data processing pipeline. Each step is discussed in a separate section.

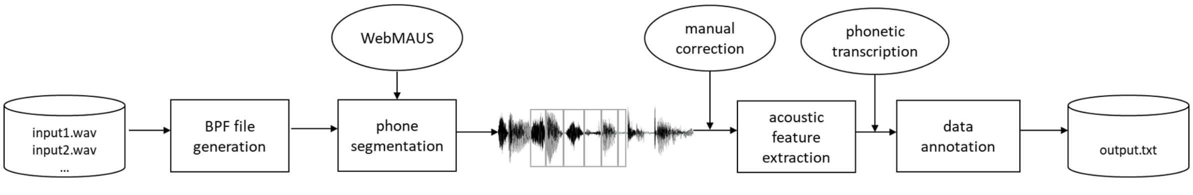

Figure 6.1 Pipeline of data processing for acoustic analysis

6.3.1 BPF file generation

Firstly, BAS Partitur format (BPF) files were generated for automatic phone segmentation (see Section 6.3.2) based on each audio file from the word list. The Partitur format was developed by Schiel and colleagues to be used for Bavarian Archive for Speech Signals (BAS) applications (Schiel et al., 1998). In this study, Partitur file version 1.2 was adopted. The sample rate 48 kHz was also specified in the BPF file. The canonical pronunciation (KAN) tier denotes the phonemic transcription of the word list in SAMPA. For each sentence in the word list, there were three words. Therefore, the KAN tier in the BPF file had three SAMPA transcriptions (KAN 0, 1, 2). Below is an example of a BPF file of *Say figzit again* from the word list:

Example:

```
LHD: Partitur 1.2
SAM: 48000
LBD:
KAN: 0 s eI
KAN: 1 f I g z I t
KAN: 2 @ g e n
```

6.3.2 Semi-automatic segmentation

In this pipeline, a semi-automatic segmentation approach was adopted. Semi-automatic segmentation generally refers to "the process whereby automatic segmentation is followed by manual checking and editing of the segment boundaries" (Gibbon et al., 1997, p. 153). As the current study examines the fricatives of Hong Kong English using pseudo-words, such kind of training data for the state-of-the-art aligners is sparse and the output boundaries are not entirely accurate. Therefore, all the output boundaries from the automatic phone segmentation were manually checked and corrected.

Automatic phone segmentation

Automatic phone segmentation was performed using WebMAUS, which is the web service for the Munich Automatic Segmentation System (MAUS) (Kisler et al., 2017; Schiel, 1999). It uses forced alignment, which is a technique developed for automatic speech recognition systems. It compares "the observed speech signal and the pre-trained Hidden Markov Model (HMM) based acoustic models" (Yuan, Lai, et al., 2018, p. 1). In general, MAUS involves the following processing steps. First, the input orthographic texts are processed using a toolkit called Balloon (Reichel, 2012), which involves text normalisation and part-of-speech tagging. Examples of text normalisation include spelling out the numbers and expanding abbreviations. It is followed by tokenising the final word chain (Schiel, n.d.), meaning the word chain is split into smaller components (e.g. words and morphemes). Second, the phonemic or canonical transcriptions of the tokenised texts are generated using the grapheme-to-phoneme algorithm, with a set of probabilistic pronunciation rules applied. That is to say, not only the phonemes but also their variants are generated (where applicable). For example, in the standard British English (GB) model in MAUS, the deletion of the plosive is generated as a variant of the plosive in word-final position (@,d,#>@,#), or the voiced fricative is generated as a variant of the voiceless fricative when preceded by a vowel (@,s,#>@,z,#). Finally, the input speech signals are time-aligned to the selected phone. The selection is based on *a priori* statistical weight of each phone and the acoustic probabilities from the acoustic model.

The processing steps of MAUS were optimised to suit the aims of the present study. Commonly, the inputs for the aligner are a paired audio file and an orthographic transcription. The words are then mapped into (a grid of) possible phone sequences by using a pronunciation dictionary and/or grapheme-to-phoneme rules (Reichel, 2012; Yuan, Lai, et al., 2018). In MAUS, the prediction is computed using a data-driven decision tree (Reichel et al., 2008), which is trained on real corpus data. To improve the decision process, the results of part-of-speech tagging and morphological segmentation are also considered. However, this study adopted pseudo-words and the phoneme clusters might be different from real English words. The part-of-speech tagging might also be erroneous and create noise for the decision process. Hence, the grapheme-to-phoneme predictions might be less accurate. Moreover, the calculation of phone weights and phones to speech signal path relies on the specified acoustic model. However, the acoustic model of Hong Kong English is not available in MAUS. As a result, some optimizations were implemented. The grapheme-to-phoneme prediction was not adopted. Instead, the phonemes were directly defined by the researcher. As can be seen in the BAF file (see Section 6.3.1), there was no orthographic tier but only the canonical phone tier (KAN). "Forced alignment to input SAMPA transcript" was implemented, which left only one phoneme option for MAUS without generating any variants. The standard British

English (GB) MAUS was chosen as the acoustic model since British English has been the norm in English language teaching in Hong Kong. The GB MAUS model was trained on the AIX-MARSEC corpus (Auran et al., 2004), which contained 55,000 transcribed words of spoken standard British English. In GB MAUS, there are 165 additional pronunciation rules. None of them were applied to ensure the output label of phone was the same as in the input SAMPA transcript.

22,032 phones were from the target pseudo-words were automatically segmented and labelled, among which, 6912 were the target fricatives.

Manual correction

The output of automatic phone segmentation was a TextGrid file which contained information of the tier and time interval of each defined phone. The output boundaries of the phones from the pseudo-words (n = 22,032) were manually reviewed and corrected in Praat (Boersma and Weenink, 2018) by the researcher and a trained student helper independently. The manual correction followed the conventions for segmentation with respect to the oscillogram (Ellbogen, 2006; Jongman et al., 2000):

Fricative onset: the start of the increase in amplitude of noise
Fricative offset: the end of the decrease in amplitude of noise
Plosive onset: the start of occlusion, which is about 20-40 ms before the burst
Plosive offset: the end of the decay of aspiration
Vowel onset: the start of the periodic wave
Vowel offset: the end of the periodic wave

Apart from the oscillogram, the spectrogram was also taken into account. Given that fricatives usually have a higher frequency, for a clearer visual identification of the darkened area (i.e. high frequency of noise) in the sound spectrum, frequencies lower than 750 Hz were filtered. The window length was set to 25 ms. Figure 6.2 is an example of the oscillogram, the spectrogram, and the segmented tier after manual correction of the pseudo-word *figzit*.

Inter-rater reliability (or agreement) was calculated for the onset and offset boundaries respectively. 20 ms tolerance is commonly used when comparing segmentation results (see Yuan, Ryant, et al., 2013; Hosom, 2009). In this study, any onset and offset difference less than 20 ms was considered acceptable. Since the agreement did not involve random categorical guessing from the two segmenters (i.e. no chance agreement), a simple measure using percent agreement between segmenters was adopted.

Table 6.5 Agreement percentages within 20 ms tolerance

	onset	offset	total
overall (n=22,032)	85.5	83.9	84.7
fricatives (n=6912)	91.5	93.8	92.6

Overall, 85.5% of agreement were achieved for all phone boundaries, and 92.6% of agreement were achieved for the target fricatives (see Table 6.5 for details). The boundaries which

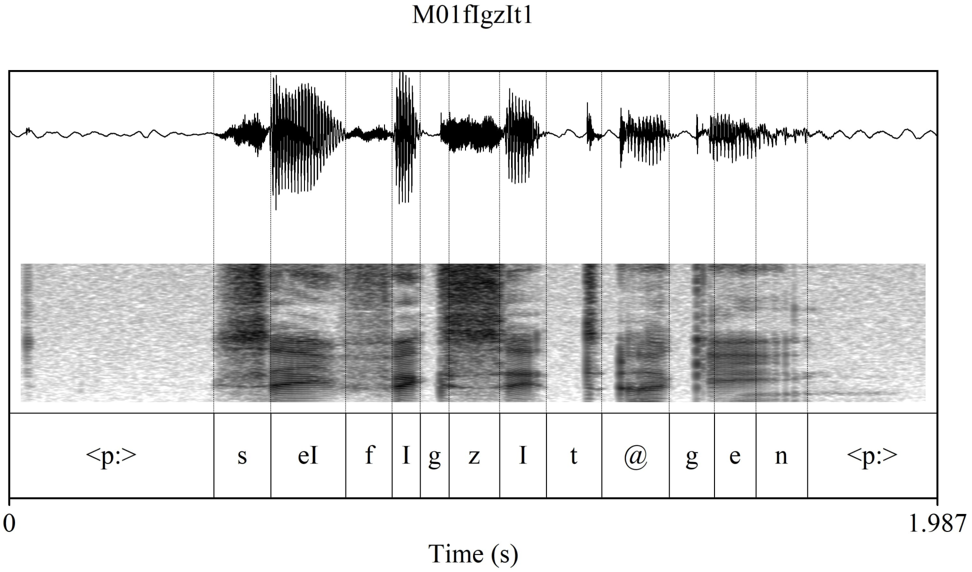

Figure 6.2 Oscillogram, spectrogram, and the phone segmentation tier of the pseudo-word *figzit* in Praat (M01figzit1)

exceeded the 20 ms range were re-examined by the researcher. Changes were made when necessary. The corrected boundaries were then input to the pipeline to extract the acoustic features.

6.3.3 Acoustic feature extraction

After the semi-automatic segmentation, acoustic measurements of the target fricatives were extracted using a Praat script written by the researcher. Since the extraction of measurements did not require any visualization of oscillogram or spectrogram, the usual Graphical User Interface (GUI) of Praat was not necessary in this case. The Praat script was called in a shell/bash script. The acoustic feature extraction procedure was embedded in the pipeline and run in the Terminal directly.

All the audio files, as a Sound object in Praat, were first downsampled to 24 kHz to reduce the data size, and hence, reduce the processing time. Resampling the input signals to 24 kHz resulted in a new Nyquist frequency of 12 kHz which should be sufficient to preserve the signal information of fricatives. Downsampling may violate the Nyquist-Shannon Sampling Theorem as the new sample rate may be less than twice the signal's bandwidth (or sampling frequency) and produce aliasing. Aliasing is an undesirable phenomenon of "the ambiguity of a sampled signal" (Boersma and Weenink, 2004, p. 884). Therefore, anti-aliasing low-pass filtering was also performed prior to resampling. The signals were then further bandpass-filtered into a number of Hann frequency bands (Boersma and Weenink, 2004) with the upper edge at 11 kHz. This removed extraneous high frequency energy, which was not relevant to the present study (Stuart-Smith, 2020). Other studies also adopted a similar range of low-pass filter (e.g. 11 kHz in the study by Jongman et al. (2000); 10,313 Hz in the study by Bukmaier and Harrington (2016); 12 kHz in the study by Jannedy and Weirich (2017)). Unless otherwise specified, the acoustic features were measured across the central 80% of the duration of the fricatives to avoid co-articulation effects.

A 25 ms Hamming window was adopted and the measurements were taken at three locations: onset (25%), middle (50%), and offset (75%). Means of the three windows were computed and reported. To measure the spectral moments, the Sound object was first converted into a Spectrum object using the fast Fourier Transform algorithm. The fast version of Fourier Transform involves padding zeroes to the sound until the number of samples N is the next highest power of two such that "the computation time scales as N log N" (Boersma and Weenink, 2004, p. 740). The FFT spectrum was used to compute the four spectral moments.

Some data removal procedures were taken before extracting the acoustic features. In general, phones with a duration less than 50 ms were automatically excluded from extracting the acoustic features. Since the acoustic features were measured across the central 80% of the interval duration, phones which were too short would cause a runtime error.

In the rest of this section, the algorithms of measuring the acoustic characteristics, the commands, and the values of the parameters used in the Praat script are introduced. How the acoustic features were described with respect to fricatives in the literature is discussed in Section 3.2.

Centre of gravity (CoG)

The centre of gravity of a spectrum is a measure of the averaged energy concentration or how high the frequencies on average are located in a spectrum (Boersma and Weenink, 2004). It is the average frequency across the entire frequency domain, weighted by energy. In mathematical terms, it can be calculated by dividing the weighted complex FFT spectrum $S(f)$ with the frequency f by energy, as shown in Equation 6.1. In the script, the query command *Get centre of gravity* was used and the parameter p was set to two, meaning the weighting was done by the power spectrum but not the absolute spectrum.

$$\frac{\int_0^\infty f|S(f)|^p df}{\int_0^\infty |S(f)|^p df} \tag{6.1}$$

Standard deviation (SD)

The standard deviation of a spectrum is a measure of "how much the frequencies in a spectrum can deviate from the centre of gravity" (Boersma and Weenink, 2004, p. 778). In Praat, it is defined as "the square root of the second central moment of [the] spectrum" (Boersma and Weenink, 2004, p. 778). Here, central moment is calculated using the Equation 6.2.

$$\frac{\int_0^{\infty}(f-f_c)^n|S(f)|^p df}{\int_0^{\infty}|S(f)|^p df} \tag{6.2}$$

$S(f)$ is the the weighted complex FFT spectrum, f is the frequency, and f_c is the centre of gravity. When the variable n is set to two, the variance of the frequencies in the spectrum is computed, which is also known as the second spectral moment. Taking the square root of it is the standard deviation of the frequency. The spectral standard deviation was extracted using the query command: *Get standard deviation*, and p was set to two in the script.

Skewness

The skewness of a spectrum reflects the spectral tilt, which is "the overall slant of the energy distribution" (Jongman et al., 2000, p. 1253). It is a measure of "how much the shape of the spectrum below the centre of gravity is different from the shape above the mean frequency" (Boersma and Weenink, 2004, p. 779). In Equation 6.2, when n is set to three, the non-normalised spectral skewness is calculated. Normalised skewness is calculated by dividing the non-normalised skewness by 1.5 power of the non-normalised second moment (variance). The normalised skewness was extracted using the query command: *Get skewness*, and p was set to two in the script.

Kurtosis

The kurtosis of a spectrum concerns the tails of the distribution. In Equation 6.2, when n is set to four, the non-normalised spectral kurtosis, which is also called the fourth spectral moment, is computed. The normalised kurtosis is calculated by dividing the non-normalised kurtosis by the square of the second moment (variance) and subtract three. The normalised kurtosis was extracted using the query command: *Get kurtosis*, and p was set to two in the script.

Peak

The spectral peak is a measure of the highest location of the shape of the spectrum, which is also the frequency "associated with the maximum energy density" (Boersma and Weenink, 2004, p. 1132). In the script, a pre-emphasis filter was applied which the spectral slope above 100 Hz increased by 6 dB/octave. The reason for using a pre-emphasis filter is that "the pre-emphasis creates a flatter spectrum, which is better for formant analysis" (Boersma and Weenink, 2004, p.191). Boersma and Weenink (2004) also stated that "we want our formants to match the local peaks, not the global spectral slope" (p. 191). The Sound object was then transformed to long-term average spectrum (Ltas), which "represents the logarithmic power spectral density as a function of frequency" (Boersma and Weenink, 2004, p. 631). When predicting the spectral shape (curve) for Ltas object, the round to sample interpolation method was selected. It basically took the greatest available value. The peak value was extracted using the query commands: *To Ltas*, *Get frequency of maximum* in the script.

Slope

The spectral slope is a "basic approximation of the spectrum shape by a linear regression line" (Mitrović et al., 2010, p. 114). As described in Section 3.2, there are different interpretations of a spectral slope. This study measured the low-frequency slope, which Stuart-Smith (2020) described as "the low frequency 'shoulder' of energy visible on the spectrogram below the main bands of high frequency energy" (p. 3). Similar to measuring the spectral peak, a pre-emphasis filter was applied which the spectral slope above 100 Hz increased by 6 dB/octave. The Sound object was then transformed to Ltas object. Slope was extracted using the query command: *Get slope* over the range 1000 Hz to 4000 Hz. The same range was also used in the study by Jesus and Shadle (2002) and Stuart-Smith (2020).

F2 Onset

Formants are bands of energy resulting from the resonance of vocal tract. It is generally acknowledged that the the first formant (F1) is inversely related to the tongue height and the second formant (F2) is related to the tongue frontness/backness. Previous studies showed that the frequency transition of F2 can predict the place of articulation of different fricatives (Jongman et al., 2000). The onset F2 frequencies were estimated using the first 25% of the following vowel of the fricatives in the syllable onset position. A Formant object was created from the Sound object in Praat using the Burg algorithm (*To Formant (burg)*) (Boersma and Weenink, 2004). The maximum frequency search range was set to 5500 Hz and the frequencies above 50 Hz were pre-emphasised by a slope of +6 dB. The mean F2 values (in Hz) were extracted using the query command *Get mean: 2* from the Formant object.

Harmonics-to-noise ratio (HNR)

The Harmonics-to-Noise Ratio, which is also called Harmonicity, is a measure of the degree of acoustic periodicity (Boersma and Weenink, 2004). In the script, a Harmonicity object was first created using an algorithm described in the study by Boersma (1993), which "performs an acoustic periodicity detection on the basis of a foward cross-correlation analysis" (Boersma and Weenink, 2004, p. 571). In Praat, "Harmonicity is expressed in dB: if 99% of the energy of the signal is in the periodic part, and 1% is noise, the HNR is $10 * log10(99/1) = 20$ dB. A HNR of 0 dB means that there is equal energy in the harmonics and in the noise" (Boersma and Weenink, 2004, p. 516). Since it uses a log function, a negative value of Harmonicity indicates that there is more noise than periodicity in the signal. The maximum value is +/-20 dB. A 10 ms time step and a minimum pitch of 60 Hz was set respectively. The mean harmonics-to-noise ration was extracted using the query command *Get mean* from the Harmonicity object.

Normalised amplitude

Although the root-mean-square value of the selected Sound object can be computed in Praat, what is actually computed is the sound pressure expressed in Pascal. To extract the value in decibel (dB), the *intensity* function in Praat was used. It is equivalent to the logarithm of the root-mean-square values times 20. The term amplitude is employed hereafter. The amplitude of the fricative amplitude was then subtracted from the neighbouring vowel amplitude in the pseudo-word. By doing so, the amplitude differences among speakers can be normalised (Jongman et al., 2000).

Normalised duration

The absolute duration was measured across the whole segment. Since the absolute duration "may vary as a function of speaking rate", the normalised duration was measured as "the ratio of fricative duration over word duration" (Jongman et al., 2000, p. 1259).

Discrete cosine transform (DCT) coefficients

The discrete cosine transform (DCT) coefficients were computed using the package emuR (Winkelmann et al., 2020) in R (R Core Team, 2021). First, the paired audio files and the TextGrid files were converted to the emuDB format. Similar to the measurements of other acoustic properties, the discrete Fourier Transform (DFT) spectrum was created using the fast Fourier Transform (FFT) algorithm. The output was a power spectrum in dB from 0 Hz to 1200 Hz, which was the Nyquist frequency after downsampling the audio files from 48 kHz to 24 kHz. A filter of 500 Hz to 11 kHz was applied, which was similar to the frequency range in (Bukmaier and Harrington, 2016; Jannedy and Weirich, 2017). The DCT coefficients were estimated using Equation 6.3 (Watson and Harrington, 1999, p. 461).

$$C(m) = \frac{2}{N} k_m \sum_{n=0}^{N-1} x(n) cos(\frac{(2n+1)(m-1)\pi}{2N}) \tag{6.3}$$

$C(m)$ is the mth DCT coefficient, and $m = 1, ..., N$; $x(n)$ is the input data, which is the trajectory of the feature being modelled; N is the length of the input data or the number of points in the trajectory; by default, k_m is set to $1/\sqrt{2}$ when $m = 1$ and set to 1 when $m \neq 1$ (Watson and Harrington, 1999). In the present study, the first four coefficients were computed for the target fricatives and their potential variants. Similar to other acoustic measurements, phone segments of which the 80% duration was less than 50 ms were excluded from the DCT analysis.

6.3.4 Phonetic transcription

The target fricatives from the subset of word list data were phonetically transcribed by the researcher and the trained student helper (who also took part in the manual correction of phone boundaries) independently. The phonetic transcription was performed with the help of the audio, the oscillogram, and spectrogram. Percentages of agreement between the two labellers were calculated, as shown in Table 6.6.

Table 6.6 Percentages of agreement of phonetic labels of the fricatives

(N=6912)	*n*	%
Overall agreement	6350	91.9
Overall disagreement	562	8.1
non-sibilants disagreement	416	6.0
sibilants disagreement	146	2.1

The overall agreement was 91.9%, which indicated a high inter-rater reliability of the labels. The disagreement (8.1%) was mainly due to non-sibilant fricatives. This was not surprising since non-sibilant fricatives were not as perceptually distinguishable as sibilant fricatives. Another disagreement came from voicing such as whether the fricative was fully devoiced (i.e. voiceless) or partially devoiced. All the disagreed tokens were checked and corrected by the researcher.

6.3.5 Data annotation

The final step in the data processing pipeline is annotation. Two sets of annotation were created, one based on the phonemic transcription and the other one based on the phonetic transcription. The following factors with respect to the fricatives were annotated for acoustic analysis:

(i) Place of articulation: labiodental, dental, alveolar, postalveolar

(ii) Voicing: voiced, voiceless

(iii) Vowel: /i, e, u, a/

Although previous studies on the acoustic characteristics of fricatives (e.g. Jongman et al., 2000; Nissen and Fox, 2005; Stuart-Smith, 2020) suggested more predictive factors such as gender and age, the present study limits the investigation to the linguistic/internal factors. Place of articulation and voicing are considered most important when it comes to the classification of fricatives. Vowel is also included as previous studies found that there is a co-articulation effect from the back rounded vowel /u/ (Johnson, 2011). It would be interesting to see if there is an effect for vowel on the acoustic characteristics of Hong Kong English fricatives. It is predicted that this co-articulation effect is not only restricted to the following vowel but also the preceding vowel. Therefore, the factor vowel for acoustic analysis refers to both preceding and following vowel. Although previous studies found main effects of gender, as well as its interaction with other factors, there were no clear hypotheses and not many discussions on the results. It makes it difficult to draw references from previous studies and make comparisons. In addition, none of the previous studies on Hong Kong English phonology (e.g. Bolton and Kwok, 1990; Hung, 2000; Deterding et al., 2008) have reported an effect of gender on the English production. Hansen Edwards (2019) included gender as one of the fixed factors in the statistical models when studying TH variation in Hong Kong English and did not find an effect. Therefore, gender was excluded from the acoustic analysis for the present study.

As for the auditory (phonetic) analysis, the following factors were annotated based on previous research on Hong Kong English phonology (e.g. Bolton and Kwok, 1990; Hung, 2000; Hansen Edwards, 2019) (see also Table 2.3):

(i) Syllable position: onset, coda

(ii) Stress pattern: stressed, unstressed

(iii) Presence of preceding /u/: yes, no

(iv) Presence of following /u/: yes, no

(v) Presence of preceding labial: yes, no

The output of the data processing pipeline is a text file per participant containing the acoustic measurements, phonemic and phonetic transcription, and the annotated factors.

6.4 Auditory analysis

In total, 27,136 target fricatives were collected from the word list. 512 tokens from two speakers were excluded due to audio problems and unidentifiable pronunciations. The remaining 26,624 tokens were phonetically transcribed. As mentioned in Section 6.3.4, 6912 tokens were transcribed by the research and the trained student helper. 19,712 tokens were first initially transcribed by the classification model (see Section 6.6). Since the model was only trained to classify 11 phones, namely /f, v, θ, ð, s, z, ʃ, ʒ, d, w, ʧ/, pronunciations apart from the 11 phones as well as deletion cannot be correctly labelled. Therefore, all the tokens were then manually checked and corrected by the researcher. For the test performance of the classification model, please refer to Section 6.6.

Among the 26,624 tokens, 26,443 were labelled as the target fricatives and their variants. The remaining 181 tokens were non-target pronunciations with counts less than 28, and they were excluded from the auditory analysis. Overall, 0.06% of the data were removed. The realisations of the target tokens are displayed in Table 6.7.

6.5 Statistical analysis

6.5.1 Data description

After some data processing using the pipeline illustrated in Section 6.3, among the 6912 target fricatives, 41 tokens were too short (i.e. < 50 ms) and were removed from acoustic feature extraction. For the acoustic analysis of fricatives, only those phone segments which were phonetically labelled as fricatives were included. In total, 6111 tokens were labelled as one of the fricatives /f, v ,θ, ð, s, z, ʃ, ʒ/ and 760 were labelled as other non-fricative variants.

Table 6.7 Distributions (%) of labels and data removal

	To be analysed (N=26,443)	Removed (N=181)	Removal (%)
/f/	[f]: 3324	[s]: 1 [v]:2 [w]:1	0.09
/v/	[v]: 189 [f]: 2919 [w]: 200	[b]: 4 [p]: 2 [s]: 2 [θ]: 4 [ʧ]: 4 Ø: 4	0.60
/θ/	[θ]: 2198 [f]: 1054	[d]: 28 [ð]: 3 [s]: 13 [ʃ]: 5 [t]: 18 Ø: 9	2.28
/ð/	[ð]: 119 [θ]: 1710 [d]: 750 [f]: 724	[b]: 5 [s]: 4 [ʃ]: 1 [t]: 9 Ø: 6	0.75
/s/	[s]: 3248 [ʃ]: 73	[f]: 4 [ʧ]: 1 Ø: 2	0.21
/z/	[z]: 373 [s]: 2868 [ʃ]: 76	[t]: 6 [f]: 2 [ts]: 2 Ø: 1	0.33
/ʃ/	[ʃ]: 3237 [s]: 60	[ʧ]: 28 [θ]: 1 [ʒ]: 1 [d]: 1	0.93
/ʒ/	[ʒ]: 121 [ʃ]: 1583 [ʧ]: 1481 [s]: 135	[f]: 4 [ts]: 4	0.24

Ø denotes deletion

Table 6.8 Response and explanatory variables in the acoustic analysis of this study

Response variable	**Description**	***n***
CoG	Mean centre of gravity (Hz)	6111
SD	Mean standard derivation (Hz)	6111
Skewness	Mean skewness	6111
Kurtosis	Mean kurtosis	6111
Peak	Mean spectral peak frequency (Hz)	6111
Slope	Mean spectral slope (dB)	6111
HNR	Mean Harmonics-to-Noise ratio or harmonicity (dB)	6111
F2 Onset	Mean F2 Onset frequency of the following vowel (Hz)	1566
Normalised amplitude	Intensity (dB) based on the root-mean-square Amplitude of fricative minus the following vowel	3040
Normalised duration	Ratio of fricative duration over word duration	6111
DCT coefficients	The first four DCT coefficients (k_0-k_3) of the fricatives	6111
Explanatory variable	**Description**	**Level**
Place	Place of articulation (labiodental, dental, alveolar, postalveolar)	4
Voicing	Voicing (voiced, voiceless)	2
Vowel	Vowel context /i, e, u, a/	4

6111 tokens were used to build the statistical models for most of the acoustic properties except for F2 Onset frequencies (n = 1566) and normalised amplitude (n = 3040), which were dependent on the following environment. Table 6.8 lists the investigated response variables and their predictors (explanatory variables) with the variable name, description and number of tokens for acoustic analysis. All the acoustic measurements were numerical data and all the predictive variables were categorical data.

Regarding the auditory analysis (N = 26,443), the response and explanatory variables are illustrated in Table 6.9. Since the aim of the auditory analysis is to study variation of Hong Kong English fricatives, all response and explanatory variables are categorical data

Table 6.9 Response and explanatory variables in the auditory analysis of this study

Response variable	**Description**	***n***
/f/	[f] and no variant(s)	3324
/v/	[v] and variant(s) [w], [f]	3308
/θ/	[θ] and variant(s) [f]	3252
/ð/	[ð] and variant(s) [θ], [d], [f]	3303
/s/	[s] and variant(s) [ʃ]	3321
/z/	[z] and variant(s) [s], [ʃ]	3317
/ʃ/	[ʃ] and variant(s) [s]	3297
/ʒ/	[ʒ] and variant(s) [ʃ], [ʧ], [s]	3320
Explanatory variable	**Description**	**Level**
SylPos	Syllable Position (onset, coda)	2
Stress	Stress pattern (stressed, unstressed)	2
Preceding /u/	Preceded by the high back rounded vowel /u/ (yes, no)	2
Following /u/	Followed by the high back rounded vowel /u/ (yes, no)	2
Preceding labial	Preceded by bilabial or labiodental consonant in word (yes, no)	2

with at least one potential variants. Homogeneous group with no variants (i.e. /f/) was excluded from examination. The factors syllable position (SylPos) and stress were examined for all fricatives, while preceding /u/, following /u/ were only examined for alveolar and postalveolar fricatives. Preceding labial was examined just for the voiceless dental fricatives /θ/.

6.5.2 Linear mixed effects model

To investigate the acoustic characteristics of fricatives of Hong Kong English, linear mixed effects models were built using the *lme4* package (Bates et al., 2015) in R (R Core Team, 2021). It primarily follows the approach and algorithms outlined in the study by Lindstrom and Bates (1988). In general, the linear mixed effects model is a linear regression model which takes both global and group-level effects into account. It is particularly suitable for repeated measures data, unbalanced data, missing data, and jointly dependent random effects (Lindstrom and Bates, 1988).

In the "language-as-a-fixed-effect fallacy", Clark (1973) critiqued the common approach of conducting only a subject analysis but ignoring the effects of items in many (psycho-) linguistic experiments. In repeated-measures experiments, there are often multiple subjects, each responding to multiple items, whereas the items are presented in multiple conditions. To explore the effects of different conditions, it is common to perform a subject analysis by grouping the means of the response variables by condition and by subject. Depending on the research questions, independent/paired t-tests or one-way/repeated-measures ANOVAs are then performed. However, taking the mean of the response variables per subject is a kind of data reduction, and hence, data loss. Also, the effect of each item is ignored. To explore the effects of items, an item analysis can be performed by grouping the means of the response variables by condition and by item.

Instead of running two separate analyses for subject and item, a mixed model allows combining random effects in a single model. Random effects refer to the expected but unpredictable random variation of the variable in the data. For example, subjects can be a random effect because each subject may have a different pitch, speech rate, or reaction time. Similarly, items can be a random effect because of the different features such as word frequency, word class, and word length. On the other hand, when designing an experiment or a material, certain factors are incorporated specifically to examine their effects on the response variables. These factors are called fixed effects because they are assumed to be constant for a given population.

In the present study, crossed, independent, random effects were generally assumed for subject (i.e. speaker) and item (i.e. pseudo-word), as opposed to nested random effects (Baayen et al., 2008). In other words, the initial models included a by-subject and a by-item random intercepts. The fixed factors were assumed for place of articulation, voicing, and vowel (see the explanatory variables in Table 6.8). A step-down model building approach was adopted to simplify the structure of the fixed effects and the interaction effects, with the help of the lmerTest package (Kuznetsova et al., 2017). The algorithm used is stated below (Kuznetsova et al., 2017, p. 9), where M refers to the Model:

1. Construct an ANOVA table for M, calculate F statistics and p values for each fixed-effects term.

2. Consider the highest order interaction effects in M. The effect with the highest p value (p_{eff}) is identified and a model without this effect M_{eff} is constructed.

3. Set M_{eff} to M. If p_{eff} is less than α level or if there are no more fixed-effects then stop, otherwise go to 2.

4. Model M from Step 2 is the final model selected by the algorithm.

In this study, the highest order interaction effect was place x voicing x vowel and the step-down approach was applied based on this model for each acoustic measurement. The best-fit models were then estimated using the restricted maximum likelihood (REML). Bonferroni *post-hoc* tests (same in the study by Jongman et al. (2000)) were performed on the best-fit models in order to compare each level of a factor.

6.5.3 Mixed binomial and multinomial logistic regression

As can be seen in Table 6.9, some response variables were binomial (with two levels) and some were multinomial (with three or four levels). For modelling binary response variables, they are generally transformed to one-dimensional variable (either success/present or failure/absent). Then logistic regression, which is treated as generalised linear model in R, is conducted. This approach can be extended to multinomial variables by computing multiple binomial logistic regressions. In this study, the *lme4* package (Bates et al., 2015) in R (R Core Team, 2021) was employed. Specifically, the *glmer* function was used to build the generalised linear mixed-effects models.

6.6 Classification of fricatives

The classification problem, in this study, is identified as a multiclass classification, with an underlying assumption that each sample belongs to one and only one class or label. Three classification models were built using the manually transcribed word list data. One for classifying place of articulation, one for classifying voicing, and one for classifying phones. The phones classification model aims to classify not only the Hong Kong English fricatives but also their potential variants based on previous studies (Bolton and Kwok, 1990; Hung, 2000) and the preliminary results of the auditory analysis. The potential variants are /d, w, ʧ/. In the word list, /d/ and /w/ were also embedded in the pseudo-words (see Section 6.2.2), and they were used in training the classification model as well. Altogether, there are 11 classes or labels: /f, v, θ, ð, s, z, ʃ, ʒ, d, w, ʧ/. All models are convolutional neural networks (CNNs) which used Mel-frequency cepstral coefficients (MFCCs) of the phone segments as the data. CNN can learn the features of the input data without any human supervision. Details of the neural networks and input coefficients are described in Section 6.6.1.

6.6.1 Convolutional neural networks (CNNs)

Mel-frequency cepstral coefficients (MFCCs)

The MFCCs of each target phone segment were computed. Segments of which the 80% duration was less than 50 ms were excluded. The steps to compute MFCCs are very similar to that of the DCT coefficients. The MFCCs were extracted from the DCT filterbank spectrum (in dB) in Praat. Equation 6.4 demonstrates the relation (Boersma and Weenink, 2004, p. 1162):

$$c_i = \sum_{j=1}^{N} P_j cos(i(j - 0.5)\pi/N) \tag{6.4}$$

where N is the number of (triangular bandpass) filters and P_j is the power in dB in the j^{th} filter.

Since CNNs require the input data to have the same shape, all the target phone segments were either padded or trimmed to 30 ms. MFCCs were computed with 15 ms window length and 5 ms shift size, and the output per each phone segment was a 12 x 54 matrix (12 MFCCs x 54 frames). Altogether, there were 7940 matrices as the data for training and testing the neural network. The label of each matrix was based on the phonetic (auditory) transcription instead of the phonemic transcription. The distribution of the labels is shown in Table 6.10.

Table 6.10 Distribution of the labels in the MFCCs data

(N=7940)	f	v	θ	ð	s	z	ʃ	ʒ	d	w	ʧ
n	1695	47	1281	29	1600	128	1293	38	989	442	398
%	21.3	0.6	16.1	0.4	20.2	1.6	16.3	0.5	12.5	5.6	5.0

A large imbalance in the distribution of the target classes is exhibited. For example, there were more voiceless fricatives than the voiced counterparts. Nevertheless, it also reflects how the phones are distributed in the phonological system of Hong Kong English fricatives, to a certain extent. Therefore, when training the model, the imbalanced distribution was preserved by implementing stratified sampling in the training and test data. 25% of the data were split for testing. In total, there were 5955 tokens for training and 1985 tokens for testing.

The CNN Architecture

In the present study, a similar CNN architecture described in Section 3.3 was adopted with two convolutional layers, each followed by a pooling layer, and 5 fully-connected layers. A masking layer was added before the convolutional layers so that the time steps with the padded values were skipped in all downstream layers. For each fully-connected layer except the last layer, a dropout layer was added which prevented over-fitting in the training dataset. A dropout rate of 0.25 was employed, meaning 25% of the nodes were randomly dropped out in the neural network. The model was trained for 500 epochs with a batch size of 200. A schematic diagram of the CNN architecture is illustrated in Figure 6.3.

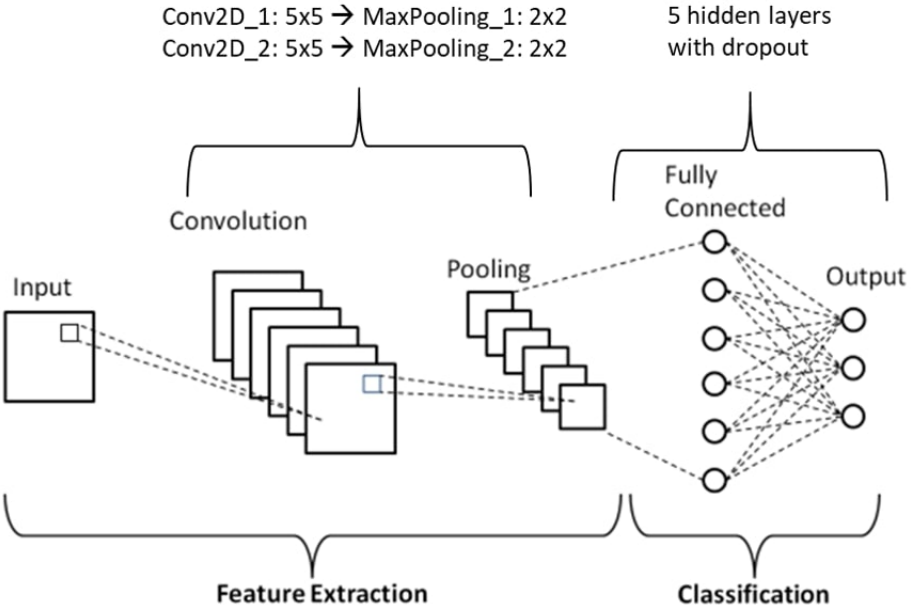

Figure 6.3 Schematic diagram of the convolutional neural network (CNN) architecture in this study, adapted from Phung and Rhee (2019, p. 3)

Specifically, the first and second convolutional layer used 32 filters with a 5 x 5 kernel size and the max pooling layer used a 2 x 2 window (strides = 2). The two-dimensional feature map was flattened to one dimensional before feeding into the fully connected layers. The fully connected network had five hidden layers with 120, 84, 60, 40, and 20 neurons respectively. The convolutional and hidden layers used the rectified linear unit (ReLU) activation function, and the output layer used the softmax activation function, which is the same in the study by Anjos et al. (2020). The softmax function is a popular function for multi-class classification and it is used in multinomial logistic regression. It normalises the input vector into a probability distribution. The output of the final layer was a list of probabilities. Therefore, a post-processing step was applied to find the highest probability and return the corresponding class label. All models were trained using Tensorflow 2 (Abadi et al., 2016).

6.7 Weighted MAUS rule set

The pronunciation rules regarding the fricative realisation in Hong Kong English were generated based on the results from the auditory analysis of the word list data with some modifications to fit the phoneme sequence in the story reading data (see Section 6.2.2). In principle, the probabilities of the realisation as different variants in the auditory analysis were computed by syllable position (onset and coda) and they were then applied to sequences in word-initial and word-final position in Table 6.4. For sequences in word-medial position, the Cambridge English Pronouncing Dictionary (Jones, 2011) was consulted to decide if the target phoneme was in the syllable onset or coda position. The same approach was adopted in the study by Hansen Edwards (2019) as well.

Although only four vowel contexts were in the word list, the rewrite patterns were applied to all vowel contexts, regardless of monophthongs or diphthongs. The same held for consonant contexts. In total, 175 rules were generated. 39 rules were about /v/, 17 were about /θ/, 31 were about /ð/, 64 were about /z/, and 24 were about /ʒ/.

As mentioned in Section 6.3.2, since the language and acoustic model of Hong Kong English was not available in MAUS, the standard British English (GB) model was employed. The GB model had its only weighted pronunciation rule set. As the rule set not only involved rewriting rules of fricatives but other groups of sound (e.g. vowels and plosives), overwriting the whole rule set might lead to drop of overall performance. Therefore, the rules generated from this study were appended to the original GB rule set. All the rules in the original GB rule set regarding fricatives were manually examined by the researcher. Original rules that were contradictory to the present study were removed, such as the rule that word-final /θ/ became /ð/ when preceded by the vowel /ɪ/ and that word-final /s/ became /z/ when preceded by a schwa /ə/.

To evaluate if the devised weighted pronunciation rules regarding fricatives actually improved the phone recognition in MAUS, two models were estimated:

(i) Standard GB MAUS modus with default rule set (Baseline GB)

(ii) Standard GB MAUS modus with generated Hong Kong English rule set (GB-HKE)

Other configurations (e.g. Pron model weight) remained the same. The phone recognition results were compared with the phonetic transcriptions conducted by the researcher. The story reading data from four participants (2F2M) were employed for the evaluation.

6.8 Summary

Figure 6.4 is a summary of the method presented in this chapter. After the data collection, the word list data and story data were subjected to different analyses. Acoustic analysis was conducted using a subset of word list data (no. of speakers = 27). Classification models were also trained on the same subset of word list data. Auditory analysis was conducted using the full set of word list data (no. of speakers = 106). Results of the auditory analysis was adopted to generate the weighted pronunciation rules of Hong Kong English fricatives. The standard British English (GB) MAUS model was adapted with the application of the Hong Kong English rule set. Finally, a subset of story data (no. of speakers = 4) was tested in order to evaluate the phone recognition performance of the adapted MAUS model.

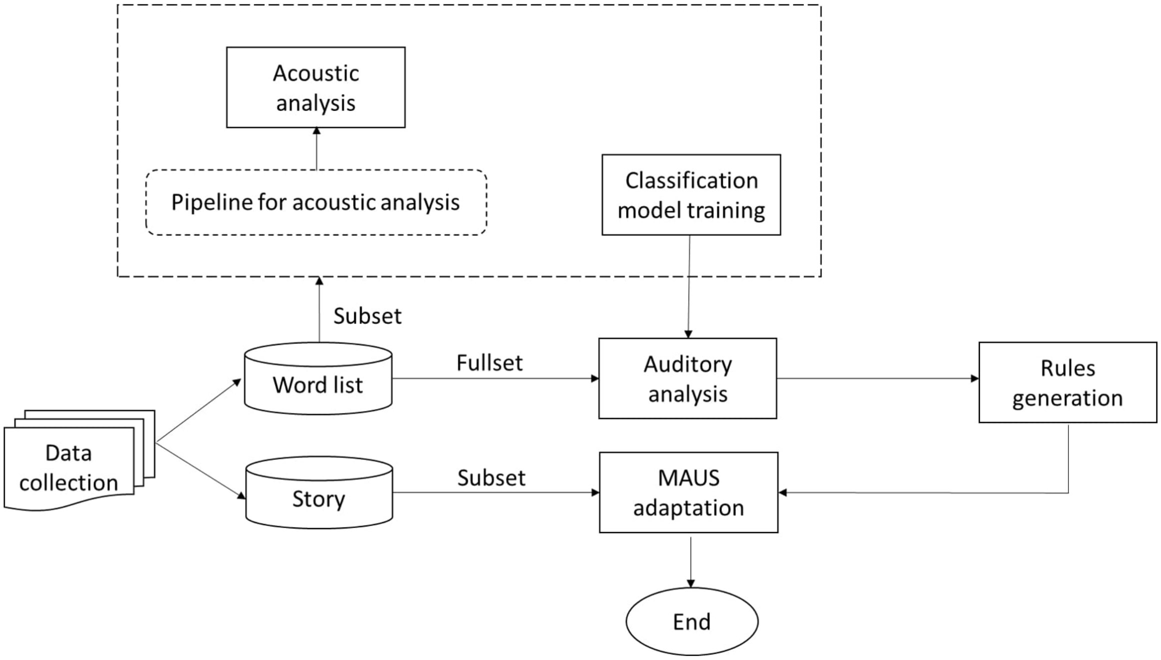

Figure 6.4 Overview of the method in the present study

Chapter 7

Acoustic analysis of Hong Kong English fricatives and their variants

7.1 Results of the acoustic analysis of Hong Kong English fricatives

The acoustic analysis was conducted using a subset (25%) of the word list data. The distribution of fricatives based on the phonetic transcription (i.e. auditory analysis) of the subset of word list data is presented in Table 7.1.

Table 7.1 Distribution of fricatives in the subset of word list data

(N=6111)	f	v	θ	ð	s	z	ʃ	ʒ
n	1695	47	1281	29	1600	128	1293	38

In the subset of word list data, all 8 English fricatives can be observed. Apart from fricatives, three major non-fricative variants, namely i) the voiced alveolar plosive [d], ii) the voiced labiovelar approximant [w], and iii) the voiceless alveolar affricate [ʧ], are also noted. The distribution of non-fricative variants is displayed in Table 7.2.

Table 7.2 Distribution of non-fricative variants in the subset of word list data

(N=643)	d	w	ʧ
n	190	55	398

7.1.1 Visualisation of fricative spectra

Before probing into specific acoustic characteristics of the fricatives in the data, the smoothed spectral shapes of [f, v, θ, ð, s, z, ʃ, ʒ] using the first four DCT coefficients (k_0-k_3) are plotted in Figure 7.1. The labels are based on the auditory analysis of the subset of the word list data. All the spectral graphs in this chapter are plotted using the package *ggplot2* (Wickham, 2016) in R. Since this package does not allow non-ASCII symbols such as some IPA symbols (θ, ð, ʃ, ʒ), all the labels in the graphs are the SAMPA symbols used in MAUS, unless stated otherwise.

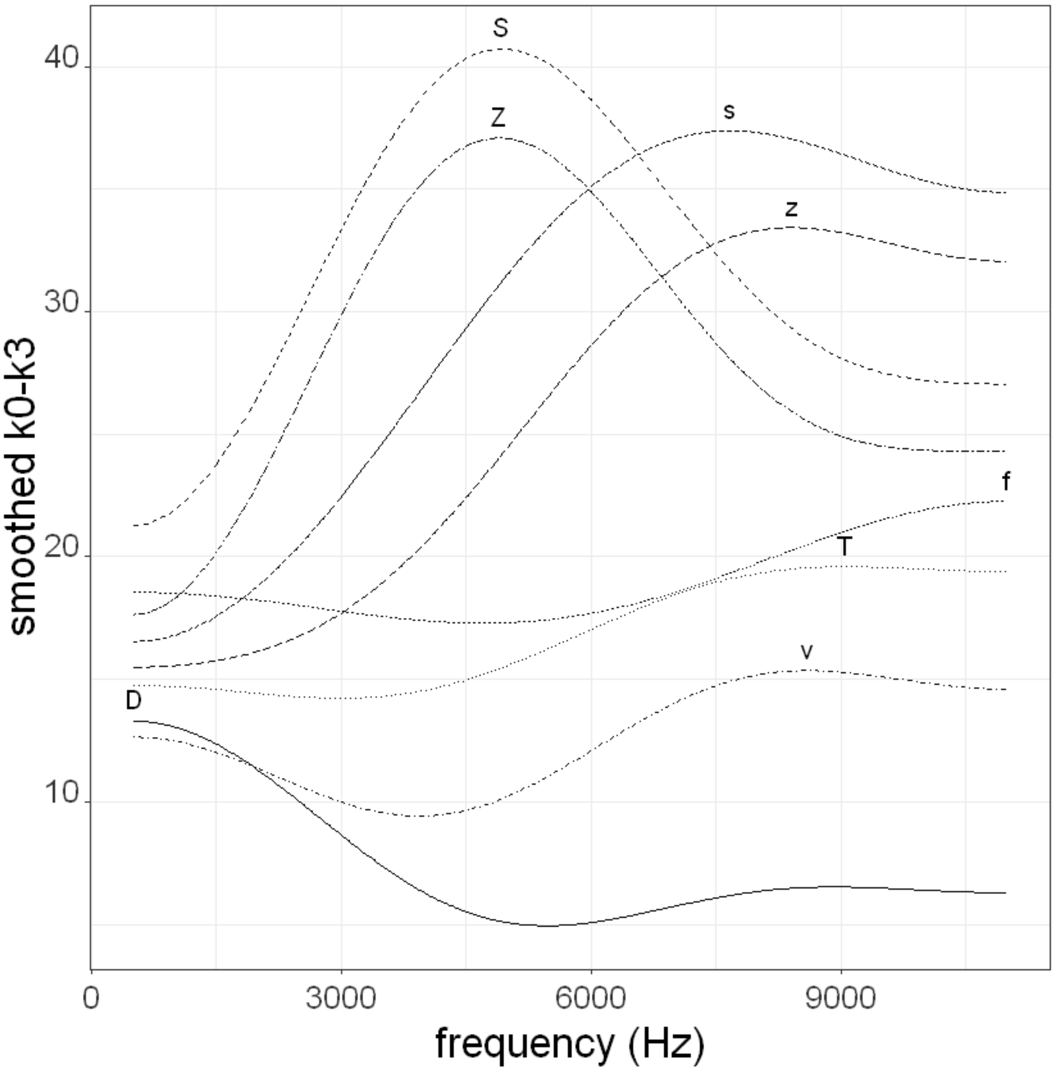

Figure 7.1 Smoothed phonetic fricative spectra using the first four DCT coefficients (k_0-k_3)

As can be seen, the smoothed spectral shapes of the labiodental, dental, alveolar, and postalveolar fricatives are distinct from each other. The overall spectra of non-sibilant fricatives are relatively flat and there are no well-defined peaks compared to the shapes of sibilant fricatives. There are differences in terms of spectral slope, curvature, and/or amplitude between the voiced and voiceless fricatives per place of articulation. For alveolar and postalveolar fricatives, it can be interpreted that the differences between the voiced and voiceless fricatives do not lie in the spectral shape but amplitude. However, it does not seem to be the case for labiodental and dental fricatives. The spectral shape of voiced dental fricative is visually different from the voiceless dental fricative, and the same holds for labiodental fricatives. That is to say, [ð] and [v] in Hong Kong English may not simply be the voiced counterpart of [θ] and [f] respectively. Therefore, in Sections 7.1.2, 7.1.3, and 7.1.4, the spectral, amplitudinal, and temporal properties of Hong Kong English fricatives are discussed respectively. With respect to the spectral slope and curvature, the two DCT coefficients (k_1 and k_2) are examined in Section 7.1.5.

7.1.2 Spectral characteristics of fricatives

Peak

Table 7.3, Table 7.4, and Table 7.5 show the mean peak values of fricatives by place of articulation, voicing, and vowel context respectively. There is an effect for place of articulation (F(3, 6069.40) = 1848.79, p < .001) on the estimation of peak values. There is also an interaction effect for place x vowel (F(9, 6069.06) = 16.11, p < .001).

Table 7.3 Mean peak values by place of articulation (n = 6111)

Place	labiodental	dental	alveolar	postalveolar
Peak (Hz)	8829.91	8472.19	7919.97	4946.24

Table 7.4 Mean peak values by voicing (n = 6111)

Voicing	voiced	voiceless
Peak (Hz)	7809.08	7508.17

Table 7.5 Mean peak values by vowel (n = 6111)

Vowel	i	e	u	a
Peak (Hz)	7759.38	7589.17	7663.68	7562.06

Bonferroni *post hoc* tests indicate that the peak differences between all places of articulation are significant (p < .001 for all pairwise comparisons), except between labiodental and dental fricatives. The estimated mean peak values of labiodental and dental fricatives alike are significantly higher than alveolar fricatives. Postalveolar fricatives have the lowest mean peak values. Regarding the interaction effect, *post hoc* tests demonstrate that the estimated peak value of labiodental fricatives is significantly higher when the vowel context is /u/.

Slope

The mean (low-frequency) slope values of fricatives by place of articulation, voicing, and vowel context are displayed in Table 7.6, Table 7.7 and Table 7.8 respectively. There is a main effect for place of articulation on the estimation of slope (F(3, 6069.44) = 4322.46, p < .001). A two-way interaction effect is also found for place x vowel on the estimation of slope (F(9, 6069.05) = 44.97, p < .001).

Bonferroni *post hoc* tests reveal that all the pairwise differences of place of articulation are significant (p < .001) except between labiodental and alveolar fricatives. The estimated mean slope of postalveolar fricatives is steepest, followed by alveolar fricatives and labiodental

Table 7.6 Mean slope values by place of articulation ($n = 6111$)

Place	labiodental	dental	alveolar	postalveolar
Slope (dB)	13.97	11.92	14.99	28.63

Table 7.7 Mean slope values by voicing ($n = 6111$)

Voicing	voiced	voiceless
Slope (dB)	16.06	17.68

Table 7.8 Mean slope values by vowel ($n = 6111$)

Vowel	i	e	u	a
Slope (dB)	16.78	16.37	18.39	16.37

fricatives alike. The estimated slope value is the smallest for dental fricatives. A two-way interaction effect is found that the estimated slope value of alveolar fricatives is significantly higher when the vowel context is /u/.

Centre of gravity (CoG)

Table 7.9, Table 7.10, and Table 7.11 show the mean centre of gravity (CoG) values by place of articulation, voicing, and vowel context. A main effect for place of articulation (F(3, 294.04) = 33.34, $p < .001$) is found. With respect to interaction effects, there is a two-way interaction effect for place x voicing (F(3, 182.71) = 6.02, $p < .001$) and place x vowel (F(9, 360.52) = 19.88, $p < .001$).

Table 7.9 Mean values of centre of gravity by place of articulation ($n = 6111$)

Place	labiodental	dental	alveolar	postalveolar
CoG (Hz)	6126.74	6309.74	7385.89	4753.15

Table 7.10 Mean values of centre of gravity by voicing ($n = 6111$)

Voicing	voiced	voiceless
CoG (Hz)	6264.75	6194.94

Table 7.11 Mean values of centre of gravity by vowel ($n = 6111$)

Vowel	i	e	u	a
CoG (Hz)	6349.57	6174.82	6286.63	6092.66

Bonferroni *post hoc* tests indicate that for place of articulation, all pairwise comparisons are significant ($p < .001$), except between labiodental and dental fricatives. Alveolar fricatives have the highest estimated mean value of centre of gravity and postalveolar fricatives have the lowest estimated value. There is an interaction effect for place x voicing: the voiced dental fricatives and the voiced labiodental fricatives have a significantly lower centre of gravity than the voiceless dental and labiodental fricative respectively. Another two-way interaction effect is found for place x vowel that the centre of gravity of alveolar fricatives is significantly lower when the vowel context is /u/.

Standard deviation (SD)

The mean values of standard deviation by place of articulation, voicing, and vowel are reported in Table 7.12, Table 7.13, and Table 7.14 respectively. A main effect is obtained for place of articulation ($F(3, 441.83) = 2003.32$, $p < .001$), voicing ($F(1, 206.28) = 9.78$, $p < .01$), and vowel ($F(3, 54.34) = 7.94$, $p < .001$) on the estimation of standard deviation. There is also a two-way interaction effect for place x vowel ($F(9, 583.49) = 20.77$, $p < .001$) on the estimated standard deviation.

Table 7.12 Mean values of standard deviation by place of articulation ($n = 6111$)

Place	labiodental	dental	alveolar	postalveolar
SD (Hz)	2799.71	2588.90	1702.96	1687.19

Table 7.13 Mean values of standard deviation by voicing ($n = 6111$)

Voicing	voiced	voiceless
SD (Hz)	2244.07	2161.53 4

Table 7.14 Mean values of standard deviation by vowel ($n = 6111$)

Vowel	i	e	u	a
SD (Hz)	2163.48	2159.89	2271.50	2198.02

Bonferroni *post hoc* tests show that all pairwise comparisons are significant for place of articulation ($p < .001$) except between labiodental and dental fricatives. Both labiodental and dental fricatives have a larger estimated standard deviation than alveolar fricatives; alveolar fricatives have a larger standard deviation than postalveolar fricatives. As for voicing, voiced fricatives have a larger estimated standard deviation than voiceless fricatives ($p < .01$). Regarding vowels, all the significant differences are due to the vowel /u/ ($p < .001$) that the standard deviation of fricatives is larger when the vowel context is /u/. *Post hoc* tests of the interaction effect for place x vowel also demonstrate that the estimated standard deviation values of alveolar and postalveolar fricatives are larger when the vowel context is /u/ ($p < .001$).

Skewness

The means of skewness by place of articulation, voicing, and vowel context are shown in Table 7.15, Table 7.16, and Table 7.17 respectively. A main effect is found for place of articulation ($F(3,\ 293.62) = 561.63$, $p < .001$) and vowel ($F(3,\ 52.88) = 10.06$, $p < .001$) on the estimation of skewness. There is also a two-way interaction effect for place x voicing ($F(3,\ 182.30) = 5.92$, $p < .001$).

Table 7.15 Mean values of skewness by place of articulation ($n = 6111$)

Place	labiodental	dental	alveolar	postalveolar
Skewness	-0.27	-0.41	-0.37	0.98

Table 7.16 Mean values of skewness by voicing ($n = 6111$)

Voicing	voiced	voiceless
Skewness	-0.096	-0.034

Table 7.17 Mean values of skewness by vowel ($n = 6111$)

Vowel	i	e	u	a
Skewness	-0.10	0.0016	-0.15	0.0096

Bonferroni *post hoc* tests show that the differences in skewness among all pairwise comparisons are significant ($p < .001$). In terms of magnitude, postalveolar fricatives have the largest estimated skewness. In terms of vowel, the main effect is primarily due to the back rounded vowel /u/ ($p < .001$). Voicing, per se, does not have a main effect on the predicted value of skewness but there is an interaction effect of place x voicing. The estimated skewness values of dental and labiodental fricatives are larger than the voiceless counterparts.

Kurtosis

Table 7.18, Table 7.19, and Table 7.20 illustrate the mean values of kurtosis by place of articulation, voicing, and vowel context respectively. An interaction effect is found for place x vowel is also found ($F(9,\ 297.11) = 6.30$), $p < .001$).

Table 7.18 Mean values of kurtosis by place of articulation ($n = 6111$)

Place	labiodental	dental	alveolar	postalveolar
Kurtosis	0.57	1.31	1.85	1.52

Table 7.19 Mean values of kurtosis by voicing ($n = 6111$)

Voicing	voiced	voiceless
Kurtosis	1.30	1.32

Bonferroni *post hoc* tests reveal the estimated kurtosis values of alveolar and postalveolar fricatives are significantly smaller when the vowel context is /u/.

Table 7.20 Mean values of kurtosis by vowel (n = 6111)

Vowel	i	e	u	a
Kurtosis	1.48	1.61	0.58	1.605

F2 Onset

The mean F2 onset frequencies of the following vowels by place of articulation, voicing, and vowel are displayed in Table 7.21, Table 7.22, and in Table 7.23 respectively. Since the F2 onset values are highly correlated with vowel, this factor is excluded from the estimated model. A main effect is found for place of articulation (F(3, 1172.84) = 114.55, p < .001) on the estimation of F2 onset values.

Table 7.21 Mean values of F2 onset by place of articulation (n = 1566)

Place	labiodental	dental	alveolar	postalveolar
F2 Onset (Hz)	1553.72	1690.56	1760.02	1845.65

Table 7.22 Mean values of F2 onset by voicing (n = 1566)

Voicing	voiced	voiceless
F2 Onset (Hz)	1715.13	1716.24

Table 7.23 Mean values of F2 onset by vowel (n = 1566)

Vowel	i	e	u	a
F2 Onset (Hz)	2132.77	1846.36	1627.38	1556.50

Bonferroni *post hoc* tests show that all pairwise comparisons of place of articulation are significant (p < .001) except between dental and alveolar fricatives. The estimated F2 onset value is highest for postalveolar fricatives, followed by dental and alveolar fricatives alike. The estimated value of F2 onset is lowest for labiodental fricatives.

7.1.3 Amplitudinal characteristics of fricatives

Normalised amplitude

The mean normalised amplitudes of fricatives by place of articulation, voicing, and vowel of pseudo-word are illustrated in Table 7.24, Table 7.25, and Table 7.26 respectively. A main effect of place of articulation (F(3, 1893.82) = 1195.68, p < .001) as well as voicing (F(1, 1374.75) = 80.54, p < .001) is found on the estimation of the normalised amplitude values.

Bonferroni *post hoc* tests reveal that the differences in normalised amplitude of fricatives are significant in all pairwise comparisons of place of articulation (all p < .001), except between labiodental and dental fricatives. The estimated normalised amplitude of postalveolar

Table 7.24 Mean values of normalised amplitude by place of articulation($n = 3040$)

Place	labiodental	dental	alveolar	postalveolar
Normalised amplitude (dB)	-17.57	-17.54	-6.91	-5.98

Table 7.25 Mean values of normalised amplitude by voicing ($n = 3040$)

Voicing	voiced	voiceless
Normalised amplitude (dB)	-10.90	-12.15

Table 7.26 Mean values of normalised amplitude by vowel ($n = 3040$)

Vowel	i	e	u	a
Normalised amplitude (dB)	-11.83	-11.04	-12.03	-11.41

fricatives is significantly larger than that of alveolar fricatives, and alveolar fricatives are significantly louder than labiodental and dental fricatives. The *post hoc* tests also show that voiced fricatives are significantly louder than voiceless fricatives ($p < .001$).

Harmonics-to-noise ratio (HNR)

The means of harmonics-to-noise ratio (HNR) or harmonicity by place of articulation, voicing, and vowel context are displayed in Table 7.27, Table 7.28, and Table 7.29 respectively. A main effect is found for place of articulation ($F(3, 588.80) = 534.72$, $p < .001$) and voicing ($F(1, 257.00) = 20.22$, $p < .001$).

Table 7.27 Mean values of harmonics-to-noise ratio by place of articulation ($n = 6111$)

Place	labiodental	dental	alveolar	postalveolar
HNR (dB)	3.89	5.97	1.21	-0.39

Table 7.28 Mean values of harmonics-to-noise ratio by voicing ($n = 6111$)

Voicing	voiced	voiceless
HNR (dB)	3.06	2.42

Table 7.29 Mean values of harmonics-to-noise ratio by vowel ($n = 6111$)

Vowel	i	e	u	a
HNR (dB)	2.72	2.99	2.03	3.12

Bonferroni *post hoc* tests indicate that all the pairwise comparisons of harmonics-to-voice ratio in different places of articulation are significant ($p < .001$). The dental fricatives have the largest estimated harmonics-to-noise ratio, followed by labiodental fricatives, and then by alveolar fricatives, and lastly, by postalveolar fricatives. Regarding voicing, the voiced fricatives have a significantly larger harmonics-to-noise ratio than the voiceless fricatives.

7.1.4 Temporal characteristics of fricatives

Normalised duration

Table 7.30, Table 7.31, and Table 7.32 show the mean normalised duration of fricatives by place of articulation, voicing, and vowel context respectively. The normalised duration of fricatives is the ratio of fricative duration over word duration (see Section 6.3.3). Since it is a ratio, there is no unit. A main effect for place of articulation ($F(3, 3938.43) = 148.16$, $p < .001$) is found on the estimation of normalised duration. There is no main effect for voicing, per se, but there is also an interaction effect for place x voicing ($F(3, 3617.86) = 6.18$, $p < .001$) on the estimation of normalised duration.

Table 7.30 Mean values of normalised duration by place of articulation ($n = 6111$)

Place	labiodental	dental	alveolar	postalveolar
Normalised duration (ratio)	0.1520	0.1614	0.1759	0.1869

Table 7.31 Mean values of normalised duration by voicing ($n = 6111$)

Voicing	voiced	voiceless
Normalised duration (ratio)	0.1706	0.1669

Table 7.32 Mean values of normalised duration by vowel ($n = 6111$)

Vowel	i	e	u	a
Normalised duration (ratio)	0.1817	0.1652	0.1704	0.15691

Bonferroni *post hoc* tests suggest that the differences of normalised duration among all places of articulation are significant ($p < .001$), except between labiodental and dental fricatives. The estimated normalised duration of postalveolar fricatives is significantly longer than alveolar fricative. The normalised duration of alveolar fricatives is significantly longer than labiodental and dental fricatives. Regarding interaction effects, *post hoc* tests demonstrate that the estimated normalised duration is longer for voiced labiodental and voiced postalveolar fricatives than their voiceless counterparts. Nevertheless, the estimated normalised duration of voiced alveolar fricatives is shorter than the voiceless alveolar fricatives.

7.1.5 DCT coefficients

Regarding the DCT coefficients, the mean values of k_1 of the fricatives by place of articulation, voicing and vowel context are reported in Table 7.33, Table 7.34, and Table 7.35

respectively. Results of the linear mixed models show a main effect for place of articulation (F(3, 3419.88) = 515.99, $p < .001$). Bonferroni *post hoc* tests indicate that the differences in k_1 values between alveolar and other places of articulation are significant ($p < .001$) but not between labiodental and dental fricatives as well as between dental and postalveolar fricatives. There is a three-way interaction for place x voicing x vowel on the estimation of k_1 values. The predicted value of k_1 of the voiced dental fricative /ð/ is particularly low when the vowel context is /u/.

Table 7.33 Mean values of k_1 by place of articulation (n = 6111)

Place	labiodental	dental	alveolar	postalveolar
k_1	-1.80	-2.87	-9.83	0.38

Table 7.34 Mean values of k_1 by voicing (n = 6111)

Voicing	voiced	voiceless
k_1	-5.47	-3.60

Table 7.35 Mean values of k_1 by vowel (n = 6111)

Vowel	i	e	u	a
k_1	-4.18	-3.50	-4.03	-3.19

The mean values of k_2 by place of articulation, voicing, and vowel of pseudo-word are illustrated in Table 7.36, Table 7.37, and Table 7.38 respectively. A main effect is found for place of articulation (F(3, 3430.73) = 369.05, $p < .001$) and voicing (F(1, 3443.49) = 44.07, $p < .001$). A two-way interaction effect is also found for place x voicing (F(3, 3430.72) = 27.35, $p < .001$) and place x vowel (F(9, 3424.02) = 7.25, $p < .001$) respectively.

Table 7.36 Mean values of k_2 by place of articulation (n = 6111)

Place	labiodental	dental	alveolar	postalveolar
k_2	1.46	0.13	-3.89	-7.62

Table 7.37 Mean values of k_2 duration by voicing (n = 6111)

Voicing	voiced	voiceless
k_2	-0.94	-2.78

Post-hoc tests reveal that the estimated k_2 values are all significantly different from each other (all pairwise comparisons $p < .001$). Moreover, the differences in k_2 between voiced and voiceless fricatives are also significant ($p < .05$). In terms of interaction effects, the estimated k_2 values are significantly different between voiced dental and voiceless dental fricatives, as

Table 7.38 Mean values of k_2 by vowel ($n = 6111$)

Vowel	i	e	u	a
k_2	-2.59	-2.53	-2.95	-2.54

well as for voiced alveolar and voiceless alveolar fricatives. Similarly, the estimated k_2 values are particularly low for dental and alveolar fricatives when the vowel context is /u/.

To sum up, results of the acoustic analysis in the present study indicate that the spectral, amplitudinal, and temporal measures can provide critical information about place of articulation of Hong Kong English fricatives. In fact, place of articulation is a significant predictor for all the acoustic properties being examined in this study. Table 7.39 is a summary of the key findings with respect to place of articulation.

Table 7.39 Summary of the key findings of acoustic analysis with respect to place of articulation

Property	**Predictability of place**
Spectral	
-Peak	all except labiodental and dental
-Slope	all except labiodental and alveolar
-CoG	all except labiodental and dental
-SD	all except labiodental and dental
-Skewness	all
-Kurtosis	place x vowel only
-F2 Onset	all except dental and alveolar
Amplitudinal	
-Normalised amplitude	all except labiodental and dental
-HNR	all
Temporal	
-Normalised duration	all except labiodental and dental
DCT coefficients	
-k_1	alveolar only
-k_2	all

The summary of the key findings of the acoustic analysis with respect to voicing is displayed in Table 7.40.

Table 7.40 Summary of the key findings of acoustic analysis with respect to voicing

Property	**Predictability of voicing**
Spectral	
-Peak	no
-Slope	no
-CoG	place x voicing only
-SD	yes
-Skewness	place x voicing only
-Kurtosis	no
-F2 Onset	no
Amplitudinal	
-Normalised amplitude	yes
-HNR	yes
Temporal	
-Normalised duration	place x voicing only
DCT coefficients	
-k_1	place x voicing x vowel only
-k_2	yes

7.2 Comparison of Hong Kong English fricatives and their variants

Since a number of variants are identified in the present study, an acoustic analysis of the variants of the fricatives was conducted to examine if these variants shared similar acoustic properties as the phonemes. It helps answer questions like: is the realisation of /θ/ as [f] same as /f/? The smoothed spectral shape of the variants and their comparative phones are also provided to facilitate the discussions.

Labiodental fricatives [f, v]

Figure 7.2 plotted the smoothed spectra of [f] and [v] using the first four DCT coefficients (k_0-k_3). It represents the average spectra of phonetic [f] (which is the realisation of /f/ as [f]) and [v] (which is the realisation of /v/ as [v]). The phonetic transcription is based on the auditory analysis of the subset of the word list data. As can be seen, [f] and [v] are not similar in terms of spectral shapes. Results from the linear mixed model show that there is a main effect for voicing on the estimation of k_1 value ($F(1, 79.15) = 7.79$, $p < .01$), suggesting the overall slope of [f] and [v] are significantly different from each other. There is also an effect for voicing on the estimation of centre of gravity ($F(1, 1724.70) = 71.05$, $p < .001$), skewness ($F(1, 1724.70) = 48.49$, $p < .001$), kurtosis ($F(1, 1736.60) = 8.22$, $p < .001$), peak ($F(1, 1586.10) = 21.87$, $p < .001$), and slope ($F(1, 1719.00) = 191.73$, $p < .001$). [v] has a lower centre of gravity, a larger positive skewness, a larger positive kurtosis, a lower peak value, and a lower slope value than [f].

Results from the auditory analysis of the subset of word list data show that /v/ is also

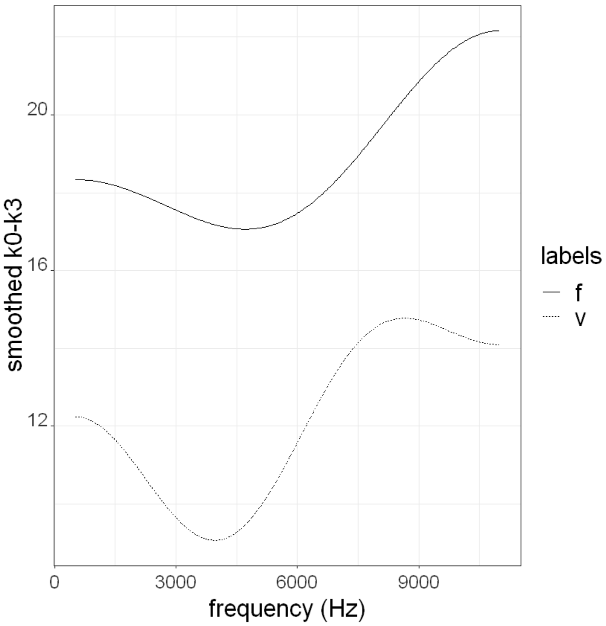

Figure 7.2 Comparison of the smoothed spectra of [f] for /f/ and [v] for /v/

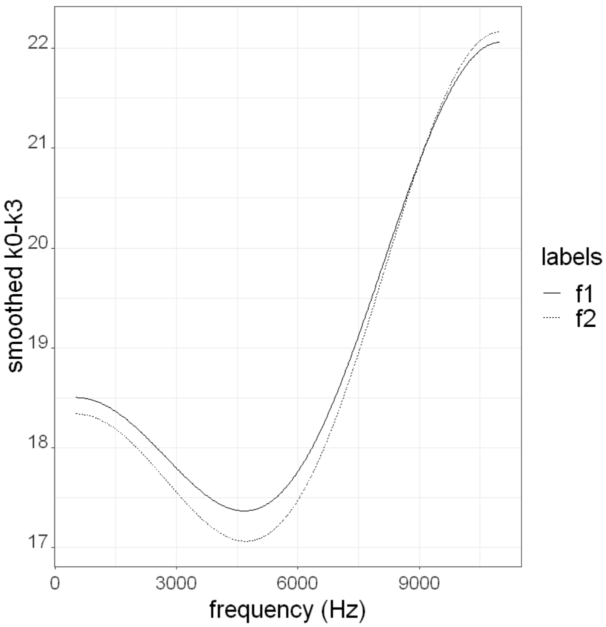

Figure 7.3 Comparison of smoothed spectra of the variant [f] of /v/ (denoted as f1) and [f] for /f/ (denoted as f2)

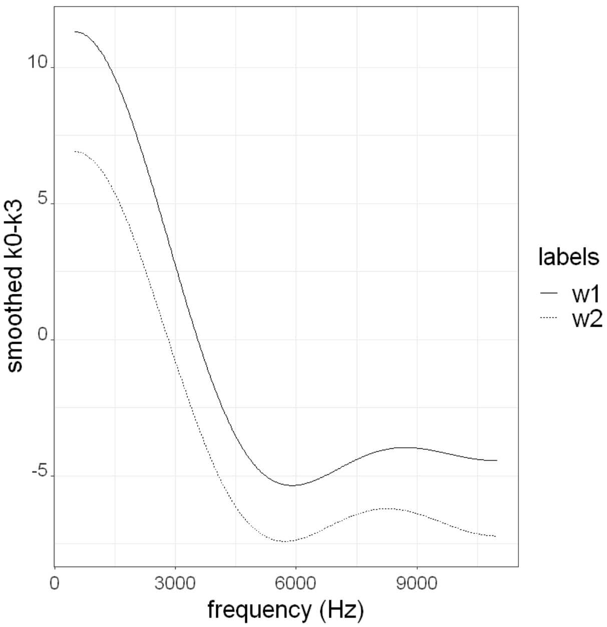

Figure 7.4 Comparison of smoothed spectra of the variant [w] of /v/ (denoted as w1) and [w] for /w/ (denoted as w2)

realised as the voiceless counterpart [f] and the labiovelar approximant [w]. The comparison of the smoothed spectral shape of the realisation of /v/ as [f] and the realisation of /f/ as

[f] as well as the comparison of the realisation of /v/ as [w] and the realisation of /w/ as [w] is plotted in Figure 7.3 and Figure 7.4 respectively. It can be seen that the averaged spectra of the variant [f] of /v/ and the [f] as in /f/ are very similar in terms of slope, curvature, and amplitude, indicating that they are very likely the same sound. The averaged spectral of the variant [w] of /v/ and the [w] as in /w/ are also similar in slope and curvature but seems to be different in amplitude. Nevertheless, results of the linear mixed model indicate that the label the normalised amplitude of the two labels are not significantly different from each other.

Dental fricatives [θ, ð]

The smoothed spectra of [θ] and [ð] are depicted in Figure 7.5. The spectral shapes of both labels seem different with respect to slope and curvature. In fact, there is a main effect for voicing on the estimation of k_1 ($F(1,\ 221.11) = 52.99$, $p < .001$) and k_2 ($F(1,\ 173.25) = 38.38$, $p < .001$). In addition, the estimated centre of gravity ($F(1,\ 29.46) = 4.20$, $p < .05$) and skewness ($F(1,\ 28.05) = 5.21$, $p < .05$) is affected by voicing.

Results from the auditory analysis of the subset of the word list data indicate that /θ/ is also realised as the voiceless labiodental fricative [f] (i.e. TH-fronting). When comparing the smoothed spectra of [θ] and the variant [f] of /θ/, as shown in Figure 7.6, the curvature of [θ] and [f] is distinct from each other. The estimated linear mixed models reveal that place of articulation is a significant predictor of k_1 ($F(1,\ 115.24) = 25.61$, $p < .001$) and k_2 ($F(1,\ 80.58) = 76.91$, $p < .001$). A main effect is also found on the kurtosis ($F(1,\ 911.25) = 5.18$, $p < .05$) that [θ] has a larger kurtosis value.

The comparison of the smoothed spectral shape of the realisation of /θ/ as [f] and the [f] as in /f/ is plotted in Figure 7.7. As can be seen, the spectral shapes of both labels are similar in terms of curvature. Although the high frequency slope of [f] for /f/ appeared steeper, the effect on k_1 is not significant. A main effect for label on the estimation of kurtosis ($F(1,\ 911.25) = 5.19$, $p < .05$) is found and the estimated kurtosis value of the variant [f] of /θ/ is significantly higher than the value of [f] for /f/. There is no effect for label on other acoustic properties.

Regarding the realisation of /ð/, it was found that the voiceless [θ] and the voiced alveolar plosive [d] (i.e. TH-stopping) are the main variants. Figure 7.8 illustrates the comparison of the variant [θ] of /ð/ and [θ] for /θ/. The spectra of both labels are very similar in shape and curvature, implying that they are the same sound. As for the realisation of /ð/ as [d], Figure 7.9 and Figure 7.10 demonstrate the spectral comparison of [ð] for /ð/ and the variant [d] of /ð/, as well as of the variant [d] of /ð/ and [d] for /d/. As can be seen in Figure 7.9, the spectral shape of [ð] and the variant [d] of /ð/ are very diverse, which is also reflected on the different estimation of k_2 ($F(1,\ 531.6) = 12$, $p < .001$). There is also a main effect for label on the estimation of kurtosis ($F(1,\ 106.58) = 8.77$, $p < .01$) that [ð] has a larger kurtosis value. With respect to the comparison of the variant [d] of /ð/ and [d] for /d/, there is a main effect for labels on the estimation of k_1 value ($F(1,\ 40.67) = 5.18$, $p < .05$) but not on k_2. No other effects are found.

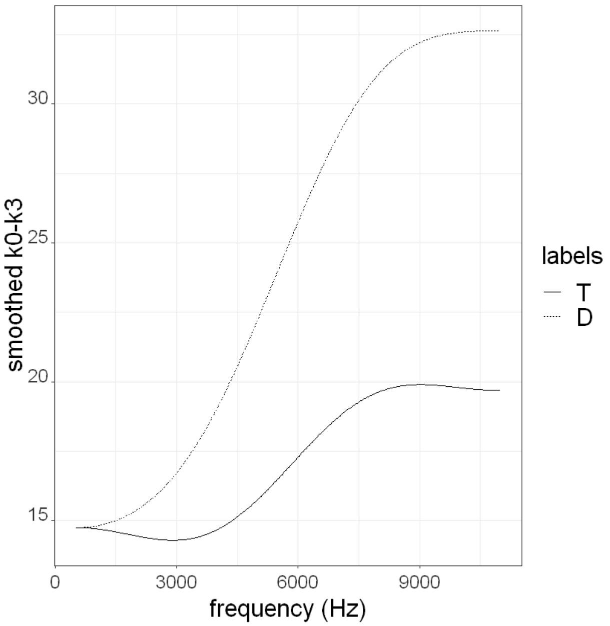

Figure 7.5 Comparison of smoothed spectra of [θ] for /θ/ (denoted as T) and [ð] for /ð/ (denoted as D)

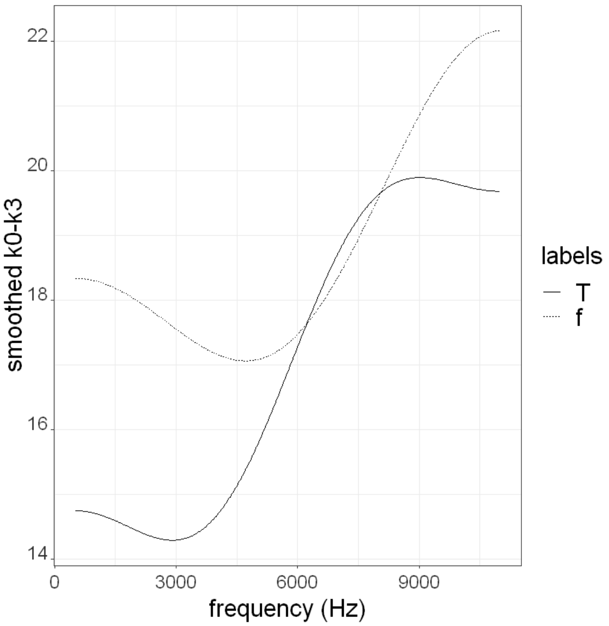

Figure 7.6 Comparison of smoothed spectra of [θ] for /θ/ (denoted as T) and the variant [f] of /θ/ (denoted as f)

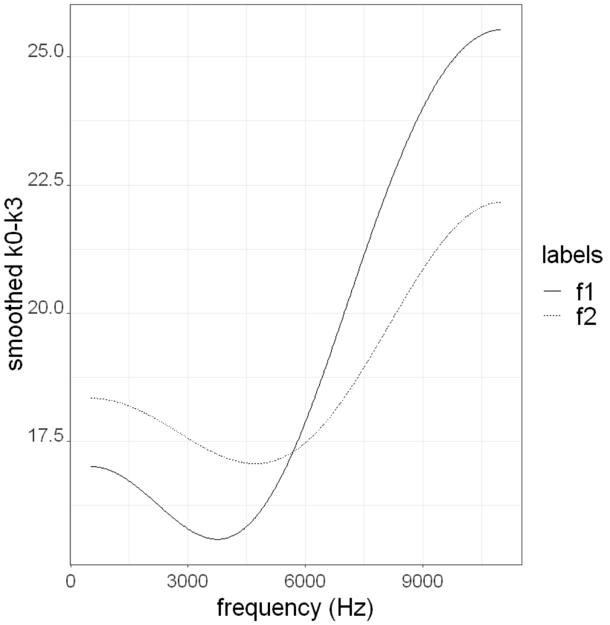

Figure 7.7 Comparison of smoothed spectra of the variant [f] of /θ/ (denoted as f1) and [f] for /f/ (denoted as f2)

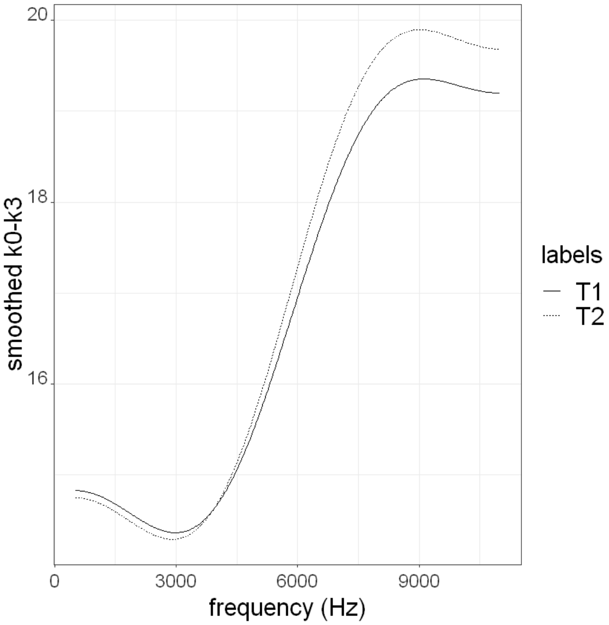

Figure 7.8 Comparison of smoothed spectra of the variant [θ] of /ð/ (denoted as T1) and [θ] for /θ/ (denoted as T2)

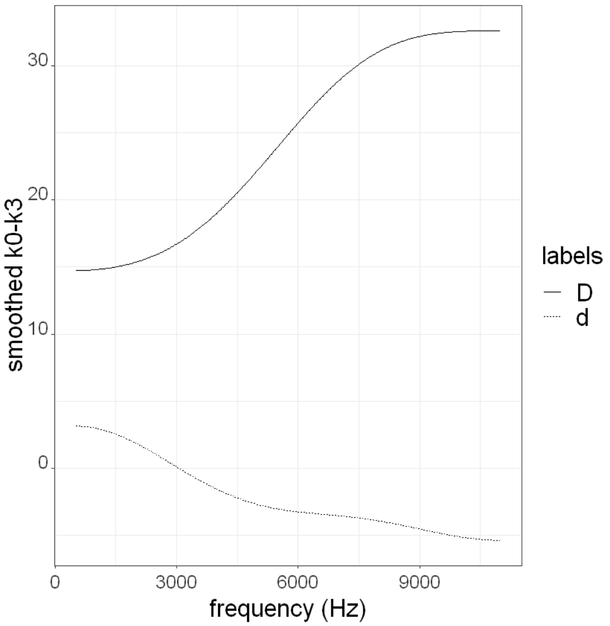

Figure 7.9 Comparison of smoothed spectra of [ð] for /ð/ (denoted as D) and the variant [d] of /d/ (denoted as d)

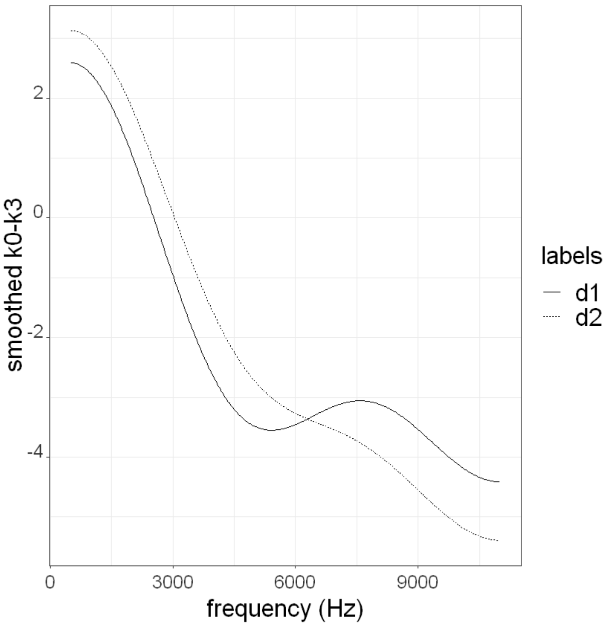

Figure 7.10 Comparison of smoothed spectra of the variant [d] of /ð/ (denoted as d1) and [d] for /d/ (denoted as d2)

Alveolar fricatives [s, z]

The comparison of the spectra of [s] and [z] is plotted in Figure 7.11. As can be seen, the spectral slope and curvature of both phones are similar. The estimated k_1 values of /s/ and /z/ are also significantly different ($F(1, 24.30) = 5.78$, $p < .05$). The estimated k_2 values are also different ($F(1, 22.68) = 5.00$, $p < .05$). The estimation of centre of gravity ($F(1, 1662.80) = 10.21$, $p < .01$) that /z/ has a larger estimated centre of gravity. Voicing also has an effect on the estimation of standard deviation is significant ($F(1, 1681.30) = 52.86$, $p < .001$) that /z/ has a larger estimated standard deviation as well as of skewness ($F(1, 1671.30) = 11.47$, $p < .001$) that /z/ has a larger negative estimation of skewness. There is also an effect on kurtosis ($F(1, 1679.40) = 31.18$, $p < .001$) that /z/ has a greater estimated kurtosis. Finally, the estimated slope of /s/ and /z/ are also different ($F(1, 1660.20) = 52.69$, $p < .001$) that /s/ has a larger positive estimation of slope.

Results from the auditory analysis of the subset of the word list data demonstrate that the main variant of the realisation of /z/ is the voiceless [s]. The comparison of the smoothed spectral shape of the variant [s] of /ʃ/ and the [s] for /s/ is plotted in Figure 7.12. The spectral shape and amplitude of both labels are very similar, and it could be concluded that they are the same phone.

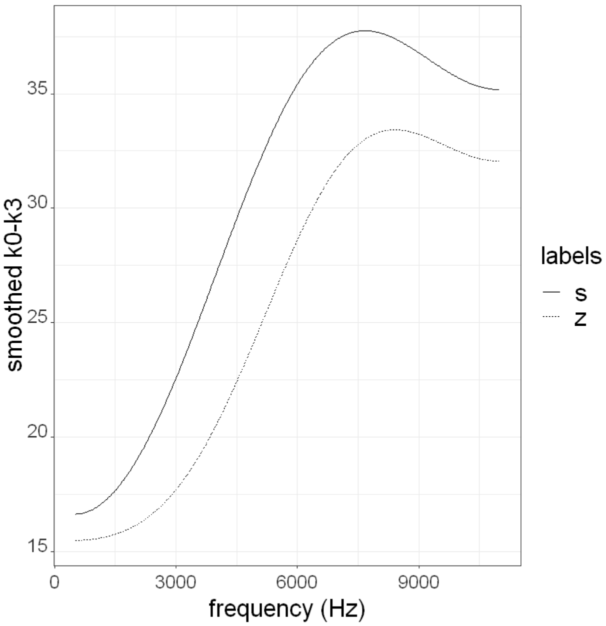

Figure 7.11 Comparison of smoothed spectra of [s] for /s/ (denoted as s) and [z] for /z/ (denoted as z)

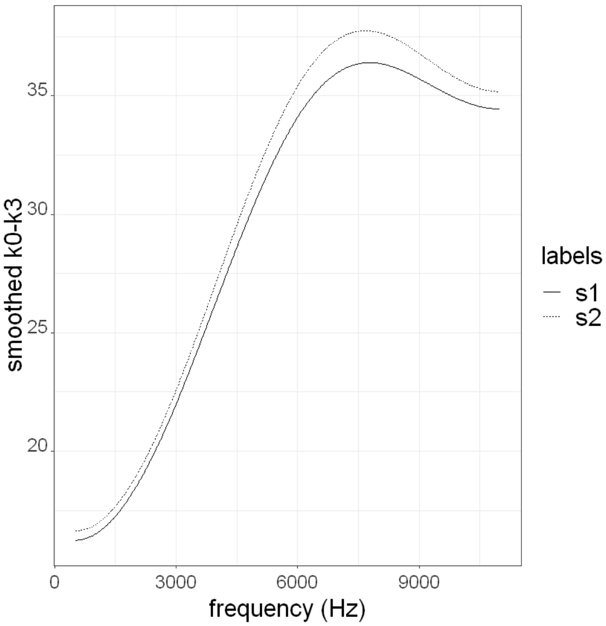

Figure 7.12 Comparison of smoothed spectra of the variant [s] of /ʃ/ (denoted as s1) and [s] for /s/ (denoted as s2)

Postalveolar fricatives [ʃ, ʒ]

A comparison of the spectral shape of [ʃ] and [ʒ] is displayed in Figure 7.13. Results from the linear mixed model show that voicing has no significant effects on the estimation of k_1 and k_2. Nevertheless, there is a main effect for voicing on the estimation of skewness (F(1, 1218.90) = 7.49, $p < .01$), as well as the estimation of slope (F(1, 1260.90) = 28.27, $p <$ .001).

As noted in the auditory analysis of the subset of word list data, the voiced postalveolar fricative /ʒ/ is also realised as the voiceless [ʃ]. Figure 7.14 illustrates the smoothed spectra of the variant [ʃ] of /ʒ/ and the [ʃ] for /ʃ/. As can be seen, the spectral shape and amplitude of both labels is almost identical, indicating that they are very likely the same phone.

In summary, for the realisations of /v/ as [f] and as [w], it was found that they are not significantly different from the /f/ and /w/. For the realisations of /θ/ as [f], it was found that there is a difference in kurtosis between the variant [f] of /θ/ and /f/. For the realisations of /ð/ as [d], it was found that the variant [d] of /ð/ has a different estimated k_1 from /d/. As for the realisations of /ʒ/ as [ʃ], it was found that the variant is not significantly different from /ʃ/. It can be concluded that the variants of Hong Kong English fricatives are not acoustically different from the respective phonemes. It is important when it comes to how many phone symbols should be established when constructing an acoustic model for Hong Kong English. For example, if it the findings of the acoustic properties of a variant were significantly different from the existing phone symbols, a new phone symbol may need to be

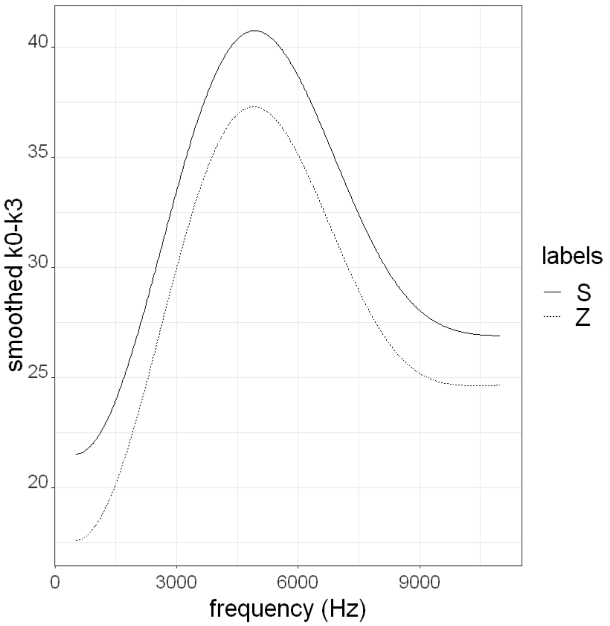

Figure 7.13 Comparison of smoothed spectra of [ʃ] for /ʃ/ (denoted as S) and [ʒ] for [ʒ] (denoted as Z)

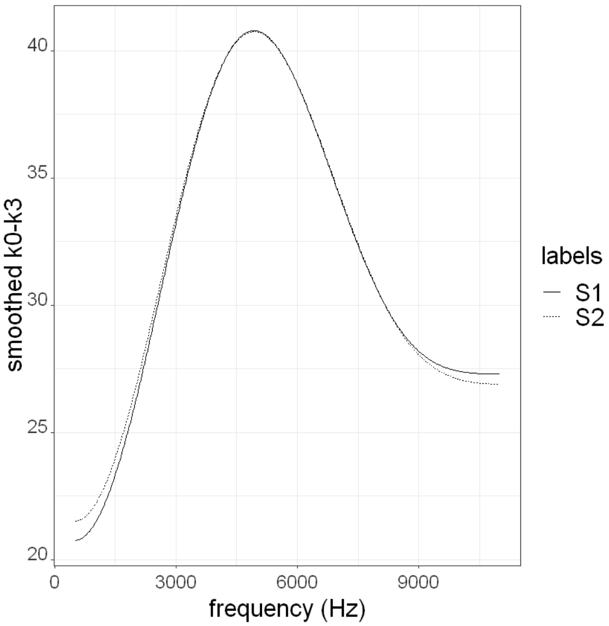

Figure 7.14 Comparison of smoothed spectra of the variant [ʃ] of /ʒ/ (denoted as S1) and [ʃ] for /ʃ/ (denoted as S2)

established. In many cases, even if a variant or a phone has unique acoustic properties, if the occurrence of that sound is not frequent in the training data, a new phone symbol may not be created. It is because the probability of backtracking that phone symbol is too low and it may never be retrieved.

7.3 Discussion on the acoustic characteristics of Hong Kong English fricatives

The following discussions attempt to answer the research questions:

(i) Which acoustic properties of Hong Kong English fricatives can distinguish all four places of articulation (i.e. labiodental, dental, alveolar, and postalveolar) and voicing (i.e. voiced and voiceless)? Are these acoustic properties for classification the same as those for Inner Circle English fricatives?

(ii) What are the acoustic characteristics of Hong Kong English fricatives? Do they share the same pattern as the Inner Circle English fricatives?

7.3.1 Place of articulation

The present study found that spectral peak, slope, centre of gravity (CoG), standard deviation (SD), F2 onset frequency, normalised amplitude, and normalised duration are likely to distinguish all places of articulation except between labiodental and dental. Only 2 out of 10 properties, namely skewness and harmonics-to-noise (HNR) ratio, are likely to distinguish all four places of articulation. This finding conforms the prediction that only a small number of properties can distinguish all four places of articulation (see Chapter 5). This finding is also in line with previous studies on the acoustic characteristics of fricatives of standard American English that only four properties can distinguish all four places of articulation. Nevertheless, the properties are not the same. For example, in the study by Jongman et al. (2000), peak, skewness, and normalised amplitude were reported to be able to distinguish all four places. In the study by Nissen and Fox (2005), standard deviation was reported to be able to distinguish all four places. Since previous studies did not examine DCT coefficients, they are excluded from comparison.

This study measured most acoustic properties at three points: 25%, 50%, and 75%, and the mean values of the property were computed. Jongman et al. (2000) measured the spectral moments at four locations (i.e. onset, middle, offset, and transition) and pointed out that it was more likely to have at least one window location which can distinguish all four places of articulation. Generally speaking, the onset and transition location carried most distinctive formation. For the onset window location, variance, skewness, and kurtosis were able to distinguish all four places of articulation. For the transition window location, spectral mean (centre of gravity), variance, and skewness were able to distinguish all four places of articulation (Jongman et al., 2000). This finding is interesting because normally the central part of the frication noise is considered to be relatively static, as demonstrated by Maniwa et al. (2009) and Stuart-Smith (2020). Therefore, many studies extracted the acoustic measurements from the central 70% or 80% of the frication noise. Since previous studies and the present study continue to the demonstrate the low success rate to distinguish labiodental fricatives from dental fricatives, perhaps comparing more window locations is a practical solution.

In terms of the pattern of acoustic characteristics with respect to place of articulation, peak shares a similar patter with previous studies that the mean value decreases as the place of articulation moves further back in the oral cavity. The mean peak value of post-alveolar fricatives is exceptionally low due to the effect of the sublingual cavity. This pattern can also be visually observed in the smoothed spectra of fricatives. Theoretically speaking, a similar effect can be caused by the vowel /u/ and an interaction effect was expected for place x vowel. Nevertheless, in this study, the interaction effect only comes from the labiodental fricatives /f, v/. As for slope, the present study measured the low-frequency slope (0 to 4000 Hz) and the findings suggest that the slopes for postalveolar fricatives are steepest, and the slopes for non-sibilant fricatives are relatively flat. This pattern can be observed visually in Figure 7.1 as well.

With respect to the spectral moments, the findings of centre of gravity follow a similar pattern as previous studies. Postalveolar fricatives have a particularly low estimated value of centre of gravity. F. Li, Munson, et al. (2011, p. 1001) explained that "the longer the front resonating cavity is, the lower the overall resonating frequencies in the fricative spectrum will be, which is reflected in a lower M1 value". It also explains why there is an interaction effect of place x vowel in the present study. Nevertheless, this effect is only limited to alveolar fricatives followed by a back rounded vowel /u/. Results of the standard deviation in the present study show that the mean values of non-sibilant fricatives are significantly higher than sibilant fricatives, which is in line with previous studies. It means that energy is more dispersed in non-sibilant fricatives. As mentioned in Section 3.1, for non-sibilant fricatives, since the sound source is close to the constriction, there is higher "acoustic impedance" (Zhao, 2010, p. 128), and hence, the energy is more dispersed.

Skewness is the only overlapping acoustic property with previous studies which can distinguish all four places of articulation. Nevertheless, the pattern of skewness of the present study is different from previous studies. For example, Jongman et al. (2000) showed mixed results of skewness that labiodental and postalveolar fricatives were left-tilted, whereas dental and alveolar fricatives were right-tilted. In the present study, labiodental, dental, and alveolar fricatives are positively tilted (right-tilted), and only postalveolar fricatives are negatively tilted (left-tilted). Negatively tilted means energy is concentrated in lower frequencies with respect to the mean. This pattern can be visually observed in Figure 7.1 as well. The finding of the present study is more line with Nissen and Fox (2005) in which only the spectra of postalveolar fricatives were negatively tilted. Regarding kurtosis, which is related to tailedness of distribution, results from the present study found that the spectrum was less tailed for alveolar and postalveolar fricatives when followed by the back rounded vowel /u/, meaning it is closer to a Gaussian distribution. In terms of pattern, results from the present study are different from previous studies. In Jongman et al. (2000) and Nissen and Fox (2005), postalveolar fricatives had the smallest estimated kurtosis value, while in the present study, labiodental fricatives have the smallest estimated kurtosis value.

The F2 onset values in the present study follows a general pattern that it increases as the place of articulation moves further back in the vocal tract, although the difference between dental and alveolar fricatives were not significant. This pattern is consistent with previous studies. In other words, the transitional F2 onset frequency is probably negatively correlated to the front-/backness of the tongue.

With respect to amplitudinal properties, Jongman et al. (2000) found that normalised root-mean-square amplitude can distinguish all four places of articulation with postalveolar > alveolar > labiodental > dental fricatives. In the present study, it is also found that postalveolar fricatives are the loudest, followed by alveolar fricatives. Labiodental and dental fricatives alike have the least normalised amplitude. Generally, it follows the principle of the mechanical model of fricative production as described in Section 3.1. As for the harmonics-to-noise (HNR) ratio, in the present study, it was found that HNR can distinguish all four places of articulation. Postalveolar fricatives have the most noise, followed by alveolar fricatives, and by labiodental fricatives. Dental fricatives have the least noise in the phone segment. Since the HNR ratios of labiodental fricatives and dental fricatives are significantly different from each other, it can be inferred that dental fricatives exhibit more stop-like behaviour than fricative-like manner.

Regarding the temporal properties, the normalised duration, which is the ratio of fricative duration over word duration, is longest for sibilant fricatives than non-sibilant fricatives.

Among the sibilant fricatives, the normalised duration of postalveolar fricatives is longer than alveolar fricatives. There is no significant difference between the normalised duration of labiodental and dental fricatives in the present study. This finding is consistent with previous studies.

7.3.2 Voicing

Peak, slope, kurtosis, and F2 Onset between voiced and voiceless fricatives are not distinguishable. Standard deviation (SD), normalised amplitude and harmonics-to-noise ratio (HNR) are able to distinguish between voiced and voiceless fricatives. This finding is different from the predictions and from previous studies on the acoustic analysis of standard American English in which most of the acoustic properties studied report a main effect for voicing.

Findings from the present study suggest that the differences of voiced and voiceless fricatives in Hong Kong English may not lie in the spectral properties, in general, but more on the amplitudinal properties. Results of the harmonics-to-noise (HNR) ratio from the present study show that voiced fricatives have a higher HNR ratio than voiceless fricatives. This finding is consistent with previous studies and is not surprising due to the nature of more vocal fold vibration. With regards to normalised amplitude, previous studies reported that the normalised amplitude of voiceless fricatives were higher than voiced fricatives (Jongman et al., 2000). The present study shows an opposite pattern that voiced fricatives have a higher normalised amplitude than voiceless fricatives. In other words, voiced fricatives are pronounced louder than voiceless fricatives in Hong Kong English.

An interaction effect of place x voicing is reported for centre of gravity (CoG), skewness, and normalised duration in the present study. The estimated centre values of centre of gravity of voiced non-sibilant fricatives are smaller than the voiceless non-sibilant fricatives. The estimated skewness values of voiced non-sibilant fricatives were larger than the voiceless counterparts. It suggests that the behaviour of /v/ and /ð/ is different from other voiced fricatives in Hong Kong English in terms of centre of gravity and skewness. As for normalised duration, it was found that voiced alveolar fricative /z/ is shorter than voiceless alveolar fricative /s/ but not other places of articulation.

Comparisons between the voiced and voiceless fricatives have been conducted per place of articulation in order to examine if there are any differences in terms of spectral properties. Generally speaking, there are four to five properties per place of articulation which can distinguish the voiced from the voiceless fricatives. That is to say, breaking down the fricatives into place of articulation helps distinguishing voicing pattern.

Chapter 8

Classification of Hong Kong English fricatives and their variants

8.1 Results of classification of Hong Kong English fricatives and their variants

Three classification models based on the acoustic features of the fricatives and their variants are built using convolutional neural network (CNN). The classification accuracy was computed by comparing the predicted labels from the classification models (i.e. hypothesis) and the annotation based on the auditory analysis of the subset of the word list data (i.e. reference). The overall accuracy is 83.4% for place of articulation, 96.1% for voicing, and 80.6% for phone symbols, as displayed in Table 8.1.

Table 8.1 Overall classification accuracy by place of articulation, voicing, and phone symbol

	Place	Voicing	Phone
Accuracy	83.4%	96.1%	80.6%

The accuracy and weighted F1 score of place of articulation, voicing, and phone symbol are calculated to evaluate the performance of each model. F1 score, which is also called balanced F-score, is the weighted average of precision and recall. Similar to accuracy, the maximum F1 score is one, which indicates the best value, and the minimum F1 score is zero. Since F1 score can better represent the results of an imbalanced dataset, which is also the case of the present study, the results of the F1 score are mainly reported and discussed.

8.1.1 Classification of place of articulation

The accuracy and F1 score by place of articulation are displayed in Table 8.2. There are five classes of place of articulation, namely labiodental, dental, alveolar, postalveolar, and labiovelar. The F1 score is 0.88 for labiodental, 0.77 for dental, 0.94 for alveolar, 0.98 for postalveolar, and 0.90 for labiovelar. Generally speaking, the class labiovelar, labiodental, alveolar, and postalveolar can be accurately labelled based on the input acoustic signals, except for the class dental.

Table 8.2 Accuracy and F1 score of the classification of place of articulation

	labiodental	dental	alveolar	postalveolar	labiovelar
Accuracy	0.79	0.63	0.88	0.96	0.82
F1 score	0.88	0.77	0.94	0.98	0.90

Table 8.3 Confusion matrix (in count) of the classification of place of articulation.

(N=1985)	labiodental	dental	alveolar	postalveolar	labiovelar
labiodental	344	85	3	0	4
dental	102	205	17	0	3
alveolar	11	52	600	10	6
postalveolar	5	3	9	415	0
labiovelar	1	3	16	0	91

The confusion matrix is displayed in Table 8.3. In the confusion matrix, the horizontal header of labels is the reference, which is the auditory analysis of the subset of the word list data, and the vertical header of labels is the hypothesis, which is the predicted label by the model. As can be seen, 102 of the phone segments are mislabelled as dental while the correct label is labiodental. For the class dental, 85 of the phone segments are mislabelled as labiodental and 52 are mislabelled as alveolar. For other places of articulation, there are no major confusions.

8.1.2 Classification of voicing

The accuracy and F1 score by voicing are displayed in Table 8.4. There are two classes, namely voiceless and voiced. The F1 score for voiceless is 0.99 and the F1 score for voiced is 0.94. That is to say, the model is able to accurately label the voicing pattern based on the input acoustic features of the phone segment.

Table 8.4 Accuracy and F1 score of the classification of voicing

	voiceless	voiced
Accuracy	0.98	0.89
F1 score	0.99	0.94

8.1.3 Classification of allophones

In the CNN, there are 11 classes of phone symbols: [f, v, θ, ð, s, z, ʃ, ʒ, d, w, ʧ]. The overall confusion matrix is illustrated in Table 8.5. Since for this particular test dataset, there are only two instances of [v] and zero instance of [ð] and [ʒ], they are excluded from further examination and discussion.

The accuracy and F1 score for each class are computed (except for [v, ð, ʒ]) and displayed in Table 8.6. Six of the eight F1 scores ([f, s, ʃ, d, w, ʧ]) are ≥ 0.90, indicating that the

Table 8.5 Confusion matrix (in count) of the classification of phone symbols.

(N=1985)	f	v	θ	ð	s	z	ʃ	ʒ	d	w	ʧ
f	365	0	50	0	5	0	0	0	3	0	1
v	6	1	1	0	0	0	0	0	0	4	0
θ	135	0	154	0	18	0	0	0	9	4	0
ð	0	0	3	0	0	0	0	0	3	1	0
s	7	0	14	0	356	12	11	0	0	0	0
z	1	0	4	0	9	17	0	0	0	1	0
ʃ	2	0	0	0	8	0	303	0	0	0	10
ʒ	0	0	0	0	0	1	6	0	0	0	2
d	3	0	3	0	0	0	2	0	227	12	0
w	5	1	3	0	0	0	0	0	12	90	0
ʧ	0	0	1	0	0	0	13	0	0	0	86

Table 8.6 Accuracy and F1 score of the classification of [f, θ, s, z, ʃ, d, w, ʧ]

	f	θ	s	z	ʃ	d	w	ʧ
Accuracy	0.86	0.48	0.89	0.53	0.94	0.92	0.81	0.86
F1 score	0.93	0.65	0.94	0.69	0.97	0.96	0.90	0.92

classification of these phones is relatively accurate. The F1 score of [θ] and [z] is 0.65 and 0.69 respectively, which is lower than other classes but still performed better than random guessing. As can be seen in Table 8.5, the confusion of [f] is mainly due to [θ] and similarly, the confusion of [θ] is primarily due to [f] and partially due to [s]. The confusion of [z] is primarily due to [s] but not vice versa. The classification results and the confusion matrix matched the disagreement of the phonetic transcription by the two labellers, as mentioned in Section 6.3.4.

This test dataset failed to present the classification of the voiced fricatives [ð, ʒ] since they are not in the test dataset. Due to the extremely low occurrences in the training dataset, not much can be concluded in terms of the classification performances of these phones. Nevertheless, the confusion matrix (Table 8.5) indicate that these phones did not cause much confusion either. Removing these three classes from training and classification also did not improve the overall accuracy. Therefore, in the classification of the remaining word list data, 11 classes instead of 8 classes are still adopted.

8.2 Discussion on the classification of Hong Kong English fricatives and their variants

In this study, 11 phone symbols were used to represent the fricatives and their variants, as shown in Table 8.7. Since building a full acoustic model of Hong Kong English is beyond the scope of the present study, a mini version of a phone classifier for the 11 phone symbols was trained instead. Each phone symbol was associated with the speech signals of the phone segment. Since the acoustic analysis demonstrated that some patterns were different from previous studies on standard American English fricatives, the acoustic features associated with the phone symbols were also different. That is to say, although the phone symbol [ð] is used in the present study, the acoustic features may not be the same as the phone symbol [ð] in standard American English or any other variety of English.

Table 8.7 Phone symbols of fricatives and their variants

f	v	s	z	θ	ð	ʃ	ʒ	d	w	ʧ

The overall classification accuracy was 80.6%, which is significantly better than random guessing (i.e. 9.09%). The F1 scores of most classes were over 0.90 except for voiceless labiodental fricative [θ] and voiced alveolar fricative [z]. Since each phone segment was treated as a discrete input, the classification model is context-independent. It is different from the actual acoustic model (such as in MAUS), which is usually context-dependent. For classifying the phones in the pseudo-words for the present study, a context-independent model is sufficient since the distribution of the preceding and following phone was controlled. Nevertheless, for real word data, the phone sequence is meaningful as the probabilities of occurrence of certain phone combination are different (also see 10.2.1).

Chapter 9

Auditory analysis of Hong Kong English fricatives and their variants

9.1 Results of the auditory analysis

9.1.1 Overall distribution of fricatives and variants

The overall distribution of fricatives and their variants in the full set of word list data for auditory analysis is plotted in Figure 9.1. Realisation 1 refers to the realisation which is same as the target fricative label. Realisation 2-4 refer to the variants of that fricative. The corresponding frequencies of the realisations are listed in Table 9.1.

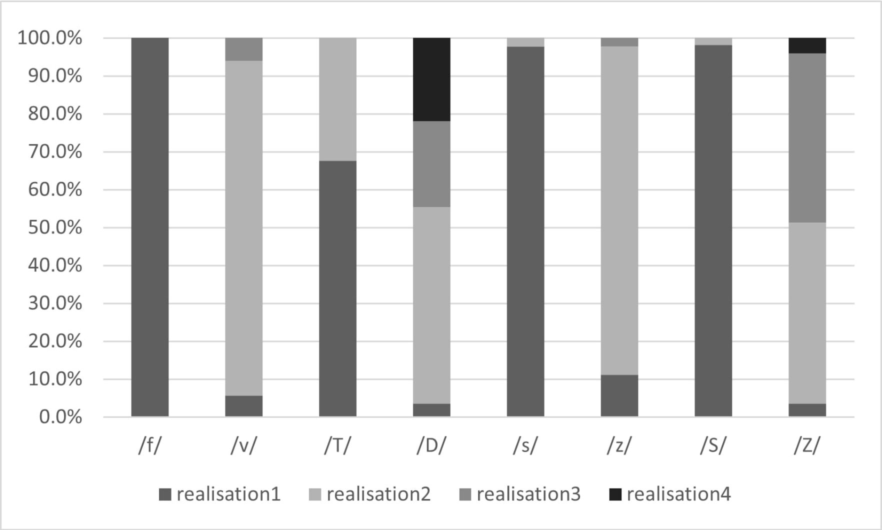

Figure 9.1 Overall distribution of fricatives and their variants in the word list dataset

As can be seen, the phonetic realisations of /f/, /s/, and /ʃ/ are relatively consistent as there is zero or only one variant, and the frequency of occurrence of that variant is less than 2.5%. /θ/ also has one variant but the variant plays a more dominant role in the realisation of /θ/ than the variant of /s/ and /ʃ/. /v/ and /z/ have two variants and one of the variants

Table 9.1 Frequencies of occurrence of the fricatives and their variants in the word list dataset.

	Realisation 1	Realisation 2	Realisation 3	Realisation 4	*n*
/f/	[f]: 3324				3324
/v/	[v]: 189	[f]: 2919	[w]: 200		3308
/θ/	[θ]: 2198	[f]: 1054			3252
/ð/	[ð]: 119	[θ]: 1710	[d]: 750	[f]: 724	3303
/s/	[s]: 3248	[ʃ]: 73			3321
/z/	[z]: 373	[s]: 2868	[ʃ]: 76		3317
/ʃ/	[ʃ]: 3237	[s]: 60			3297
/ʒ/	[ʒ]: 121	[ʃ]: 1583	[ʧ]: 1481	[s]: 135	3320

plays a major role in the phonetic realisation respectively. The realisations of /ð/ and /ʒ/ are most diverse in the dataset, as reflected in the number of variants and their proportions.

9.1.2 Labiodental fricatives /f, v/

Altogether, the number of /f/ which is realised as [f] is 3324. It comprises 100% of the /f/ in the cleaned word list dataset. Hence, it can be concluded that there are no variants of /f/ and that /f/ is unanimously pronounced as [f] in Hong Kong English.

Table 9.2 Distribution of /v/ and variant(s) by syllable position

	onset	coda	*n*	(%)
[v]	121	68	189	(5.7)
[f]	1335	1584	2919	(88.2)
[w]	200	0	200	(6.1)
n	1656	1652	3308	(100)

Table 9.3 Distribution of /v/ and variant(s) by stress pattern

	stressed	unstressed	*n*	(%)
[v]	113	76	189	(5.7)
[f]	1443	1476	2919	(88.2)
[w]	104	96	200	(6.1)
n	1660	1648	3308	(100)

The number of target /v/ is 3308, among which 189 tokens (5.7%) are realised as [v], 2919 of the tokens (88.2%) are realised as the voiceless labiodental fricative [f], and 200 of the tokens (6.1%) are realised as the voiced labiovelar approximant [w]. That is to say, /v/ is mostly realised as the voiceless counterpart [f] by the participants.

The distributions of /v/ by syllable position and stress pattern are displayed in Table 9.2 and Table 9.3. As can be seen in Table 9.2, [w] only occurs in the syllable onset position due to English phonotactics. The relation between the realisation of /v/ as [w] and the stress pattern is not significant (Estimate = -1.02, SE = -1.29, $p > .05$). As for the realisation of /v/ as [v] and the variant [f], the effects of syllable position (Estimate = -2.26, SE = 0.6361, $p < .001$) and stress (Estimate = 1.52, SE = 0.61, $p < .05$) are found to be significant. [v] is more likely to occur in syllable onset position than in coda position. [v] is also more likely to occur in stressed syllables than in unstressed syllables. There is no interaction effect between syllable position and stress pattern on the estimation of realisation of /v/.

9.1.3 Dental fricatives /θ, ð/

Regarding the voiceless dental fricative /θ/, there are 3252 tokens in the cleaned word list dataset. 2198 of the tokens (67.6%) are realised as [θ] and 1054 of the tokens (32.4%) are realised as the voiceless labiodental [f]. The distributions by syllable position, stress pattern, and presence of preceding labial are illustrated in Table 9.4, Table 9.5, and Table 9.6 respectively. Results of the mixed logistic models show that both syllable position (Estimate = 1.23, SE = 0.4772, $p < .01$) and stress pattern (Estimate 1.19, SE = 0.48, $p < .05$) have an influence on the probability of having TH variation of /θ/. The variant [f] of /θ/ is more likely to occur in syllable coda position. Moreover, [f] is more likely to occur in unstressed syllables. Nevertheless, the interaction between syllable position and stress pattern is not significant (Estimate = -0.97, SE = 0.67, $p > .05$). Presence of preceding labial consonants is not a significant predictor on the realisation of /θ/ (Estimate = -0.70, SE = 0.41, $p >$ .05).

Table 9.4 Distribution of /θ/ and variant(s) by syllable position

	onset	coda	*n*	(%)
[θ]	1118	1080	2198	(67.6)
[f]	495	559	1054	(32.4)
n	1613	1639	3252	(100)

Table 9.5 Distribution of /θ/ and variant(s) by stress pattern

	stressed	unstressed	*n*	(%)
[θ]	1116	1082	2198	(67.6)
[f]	500	554	1054	(32.4)
n	1616	1636	3252	(100)

Regarding the voiced dental fricative /ð/, altogether, there are 3303 tokens. 119 of the tokens (3.6%) are realised as [ð], 1710 of the tokens (51.8%) are realised as the voiceless dental fricative [θ], 750 of the tokens (22.7%) are realised as the voiced alveolar plosive [d], and 724 (21.9%) are realised as the labiodental fricative [f]. The distributions of /ð/ by syllable position and stress pattern are reported in Table 9.7 and Table 9.8.

Results of the estimated mixed models show that syllable position has an influence on the probability of having variation of /ð/. The variation as [θ] is more likely to occur in

Table 9.6 Distribution of /θ/ and variant(s) by presence of preceding labial consonant(s)

	present	absent	*n*	(%)
[θ]	1082	544	1626	(66.3)
[f]	554	272	826	(33.7)
n	1636	816	2452	(100)

Table 9.7 Distribution of /ð/ and variant(s) by syllable position

	onset	coda	*n*	(%)
[ð]	64	55	119	(3.6)
[θ]	702	1008	1710	(51.8)
[d]	737	13	750	(22.7)
[f]	145	579	724	(21.9)
n	1648	1655	3303	(100)

Table 9.8 Distribution of /ð/ and variant(s) by stress pattern

	stressed	unstressed	*n*	(%)
[ð]	67	52	119	(3.6)
[θ]	866	844	1710	(51.8)
[d]	384	366	750	(22.7)
[f]	336	388	724	(21.9)
n	1653	1650	3303	(100)

syllable coda position (Estimate = -2.04, SE = 0.79, $p < .001$). The variation as [d] is more likely to occur in syllable onset position (Estimate = 11.08, SE = 3.51, $p < .001$). There is no effect for syllable position on the realisation as [f]. Stress has no significant influence on the realisation of /ð/.

9.1.4 Alveolar fricatives /s, z/

There are 3321 tokens of /s/ in the cleaned word list dataset, among which 3248 (97.8%) are realised as the voiceless alveolar fricative [s] and 73 (2.2%) are realised as the voiceless postalveolar fricative [ʃ]. In other words, /s/ is primarily pronounced as [s]. The distributions of [s] and [ʃ] by syllable position, stress pattern, preceding /u/ and following /u/ are reported in Table 9.9, Table 9.10, Table 9.11, and Table 9.12 respectively. Results of the mixed logistic models reveal that both syllable position (Estimate = -1.86, SE = 1.21, $p > .05$), stress pattern (Estimate = 0.26, SE = 1.24, $p > .05$), and preceding /u/ (Estimate = 2.34, SE = 1.63, $p > .05$ have no effect on the probability of /s/ variation. The following back rounded vowel /u/ has an influence on the probability of the occurrence of the variant [ʃ] of /s/ (Estimate = 4.38, SE = 1.82, $p < .01$). [ʃ] is more likely to occur when followed by the back rounded vowel /u/.

As for the voiced alveolar fricative /z/, there are altogether 3317 tokens in the cleaned

Table 9.9 Distribution of /s/ and variant(s) by syllable position

	onset	coda	n	(%)
[s]	1636	1612	3248	(97.8)
[ʃ]	28	45	73	(22.0)
n	1664	1657	3321	(100)

Table 9.10 Distribution of /s/ and variant(s) by stress pattern

	stressed	unstressed	n	(%)
[s]	1601	1647	3248	(97.8)
[ʃ]	60	13	73	(22.0)
n	1661	1660	3321	(100)

Table 9.11 Distribution of /s/ and variant(s) by preceding /u/

	present	absent	n	(%)
[s]	380	1232	1612	(97.3)
[ʃ]	29	16	45	(2.7)
n	409	1248	1657	(100)

Table 9.12 Distribution of /s/ and variant(s) by following /u/

	present	absent	n	(%)
[s]	395	1241	1636	(98.3)
[ʃ]	21	7	28	(1.7)
n	416	1248	1664	(100)

word list dataset, of which 373 (11.2%) are realised as [z], 2868 (86.5%) are realised as the voiceless alveolar fricative [s], and 76 (2.3%) are realised as the voiceless postalveolar [ʃ]. The distributions of /z/ and its variants by syllable position, stress pattern, preceding /u/, and following /u/ are displayed in Table 9.13, Table 9.14, Table 9.15, and Table 9.16 respectively.

Table 9.13 Distribution of /z/ and variant(s) by syllable position

	onset	coda	n	(%)
[z]	208	165	373	(11.2)
[s]	1414	1454	2868	(86.5)
[ʃ]	40	36	76	(2.3)
n	1662	1655	3317	(100)

Results of the mixed logistic models demonstrate that stress has an influence on the probability of occurrence of /z/ (Estimate = -1.31, SE = 0.47, $p < .01$). [z] is more likely to occur in stressed syllables. There is no effect for syllable position (Estimate = -0.71, SE = 0.51, $p > .05$). With respect to the variant [ʃ] of /z/, it was found that preceding /u/ and stress have an interaction effect (Estimate = 20.91, SE = 6.44, $p < .01$). The variant [ʃ] of

Table 9.14 Distribution of /z/ and variants by stress pattern

	stressed	unstressed	n	(%)
[z]	225	148	373	(11.2)
[s]	1411	1457	2868	(86.5)
[ʃ]	24	52	76	(2.3)
n	1660	1657	3317	(100)

Table 9.15 Distribution of /z/ and variants by preceding /u/

	present	absent	n	(%)
[z]	36	129	165	(9.9)
[s]	351	1103	1454	(87.9)
[ʃ]	28	8	36	(2.2)
n	415	1240	1655	(100)

Table 9.16 Distribution of /z/ and variants by following /u/

	present	absent	n	(%)
[z]	46	162	208	(9.9)
[s]	344	1070	1414	(87.9)
[ʃ]	26	14	40	(2.2)
n	416	1246	1662	(100)

/z/ is more likely to occur in unstressed syllables and when preceded by /u/. The following /u/ is not a significant predictor for the occurrences of the variant [ʃ].

9.1.5 Postalveolar fricatives /ʃ, ʒ/

Altogether, there are 3297 tokens of /ʃ/ in the cleaned word list dataset, among which 3237 (98.2%) are realised as the voiceless postalveolar fricative [ʃ] and 60 (1.8%) are realised as the voiceless alveolar fricative [s]. That is to say, /ʃ/ is dominantly realised as [ʃ]. The distributions of /ʃ/ by syllable position and stress pattern are displayed in Table 9.17 and Table 9.18 respectively. Results of the mixed models show that there are no effects for syllable position (Estimate = 1.18, SE = 0.99, $p > .05$), neither for stress pattern (Estimate = -0.73, SE = 0.87, $p > .05$) on the probability of the occurrence of the variant [ʃ].

Table 9.17 Distribution of /ʃ/ and variant(s) by syllable position

	onset	coda	n	(%)
[ʃ]	1615	1622	3237	(98.2)
[s]	47	13	60	(1.8)
n	1662	1635	3297	(100)

Table 9.18 Distribution of /ʃ/ and variants by stress pattern

	stressed	unstressed	*n*	(%)
[ʃ]	1638	1599	3237	(98.2)
[s]	20	40	60	(1.8)
n	1658	1639	3297	(100)

Table 9.19 Distribution of /ʒ/ and variant(s) by syllable position

	onset	coda	*n*	(%)
[ʒ]	55	66	121	(3.6)
[ʃ]	1464	119	1583	(47.7)
[ʧ]	20	1461	1481	(44.6)
[s]	123	12	135	(4.1)
n	1662	1658	3320	(100)

Table 9.20 Distribution of /ʒ/ and variants by stress pattern

	stressed	unstressed	*n*	(%)
[ʒ]	63	58	121	(3.6)
[ʃ]	809	774	1583	(47.7)
[ʧ]	722	759	1481	(44.6)
[s]	65	70	135	(4.1)
n	1659	1661	3320	(100)

Regarding the voiced postalveolar fricative /ʒ/, there are altogether 3320 tokens in the cleaned word list dataset. 121 (3.6%) of the tokens are realised as [ʒ], 1583 (47.7%) are realised as the voiceless counterpart [ʃ], 1481 (44.6%) are realised as the voiceless postalveolar affricate [ʧ], and 135 (4.1%) are realised as the voiceless alveolar fricative [s]. In other words, the variants [ʃ] and [ʧ] are the major realisations of /ʒ/ in the dataset. The distributions of /ʒ/ by syllable position and stress pattern are reported in Table 9.19 and Table 9.20.

Results of the mixed logistic models reveal the variant [ʃ] of /ʒ/ is more likely to occur in syllable onset position syllables (Estimate = -2.13, SE = 0.82, p < .01). The variant [ʧ] 0f /ʒ/ is more likely to occur in syllable coda position (Estimate = -8.13, SE = 2.2, p < .001). The variant [s] of /ʒ/ is more likely to occur in syllable onset position (Estimate = 14.81, SE = 5.71, p < .01). No effects for stress are found.

9.2 Discussion

The following discussions attempt to answer the research questions:

(i) Which fricatives can be found in Hong Kong English and what are their distributions in terms of frequency?

(ii) Which variants of fricatives can be found in Hong Kong English and what are their distributions in terms of frequency?

(iii) Which linguistic factors (i.e. syllable position, stress, preceding labial consonants, preceding /u/, and following /u/) influence the realisation of Hong Kong English fricatives?

9.2.1 Inventory of Hong Kong English fricatives

One of the aims of the present study is to propose an inventory of Hong Kong English fricatives. It would be interesting to first look at the phonetic distribution of fricatives in the word list dataset, as plotted in Figure 9.2. Given the design of the pseudo-words in the word list, the occurrence of each fricative phoneme in the word list is ideally 3328 (12.5% of the whole dataset). Therefore, 12.5% is added as a reference line in the Figure 9.2. As can be seen, the occurrences of the voiced fricatives ([v], [ð], [z], and [ʒ]) are much lower than 12.5%. For the actual number of occurrences of each fricative, please refer to Table 9.1.

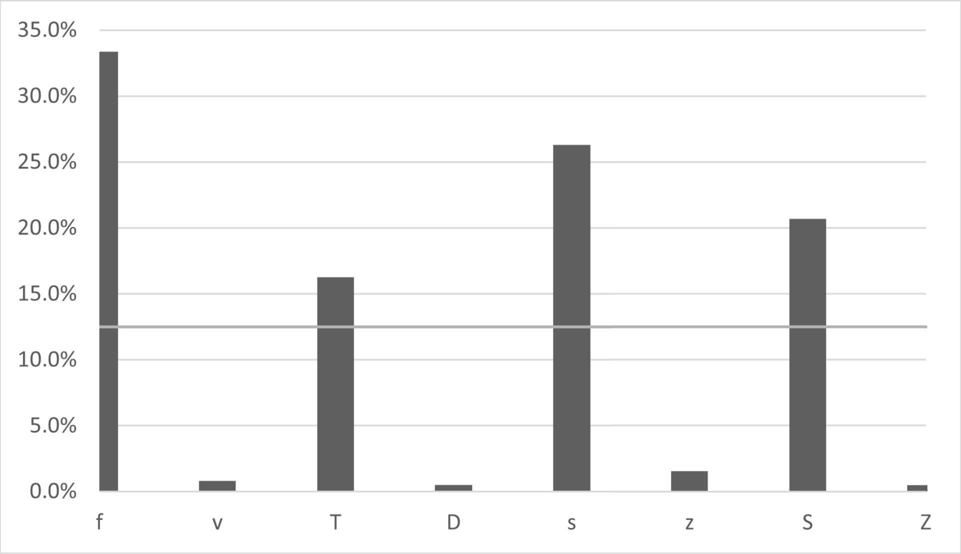

Figure 9.2 Percentage distribution of fricatives. The reference line is 12.5%.

Generally speaking, the voiced fricatives do not occur frequently in Hong Kong English. That is to say, for most mid-range Hong Kong English speakers, there are mainly four fricatives /f, θ, s, ʃ/ in their phonological system. This finding confirms the predictions stated in Chapter 5 that voiceless fricatives will be mainly found in Hong Kong English and that the occurrences of voiced fricatives will be marginal. This finding is also in line with previous studies on Hong Kong English phonology (e.g. Hung, 2000). Hung (2000) studied the speech production of 15 speakers using word list and found no evidence of [z], [ð], and [ʒ]. Deterding et al. (2008) examined the speech production of 15 speakers using interview data, and only reported that "if a sound other than [ð] is used, it is generally [d]" (p. 156) but no frequency count was provided. In the present study, 106 speakers were recruited and

26,442 tokens of fricatives were examined. Although it cannot be claimed to represent the whole population of Hong Kong English speakers, the current dataset should be generalisable enough, in which the distributions of fricative should be similar.

When it comes to the phonemic inventory of Hong Kong English fricatives, the question is always: whether there are four fricatives or eight fricatives in Hong Kong English? The present study tends to support the latter. Table 9.21 is the proposed phonemic inventory of Hong Kong English fricatives. Since the voiced fricatives are marginal, they are in brackets.

Table 9.21 Proposed phonemic inventory of Hong English fricatives

place	labiodental	dental	alveolar	postalveolar
fricative	f (v)	θ (ð)	s (z)	ʃ (ʒ)

According to Hung (2000) and Deterding (2007), the proposed phonemic inventory of Hong Kong English fricatives is different from Singapore and Malaysian English in which the voiced and voiceless contrasts of fricatives exist, although not for all places of articulation, see Table 9.22 for the inventory of consonants of Singapore English fricatives (W. Chen et al., 2010). Nevertheless, Deterding (2007) observed some instances of /θ, ð/.

Table 9.22 Phonemic inventory of Singapore English fricatives extracted from W. Chen et al. (2010)

place	labiodental	dental	alveolar	postalveolar
fricative	f v		s z	ʃ ʒ

The inventory of a variety is important when it comes to constructing the acoustic and pronunciation model of that variety. For example, W. Chen et al. (2010) used the inventory of Singapore English as stated in Table 9.22. Since there were no dental fricatives, when constructing a pronunciation dictionary of Singapore English based on the Cambridge pronunciation dictionary, the following rewrite rules were applied (W. Chen et al., 2010, p. 3):

- Dental fricative /θ, ð/ → /t,d/ in syllable-initial position
- Dental fricative /θ/ → /f/ in syllable final position

It should be noted that these rewrite rules were applied to all syllable initial or final position in W. Chen et al. (2010).

In the present study, the pronunciation rules based on the standard British English (GB) pronunciation dictionary in MAUS (see Section 6.7) are established using all the eight fricatives and their variants. In general, the voiced fricatives are set with low probabilities. If there were only four fricatives (all voiceless) in the inventory of Hong Kong English, the voiced fricatives would be completely rewritten to other sounds, as demonstrated in Singapore English (W. Chen et al., 2010), meaning the voiced fricatives would never be recognised. Nevertheless, this is not the case of the present study. The possibility to recognise voiced fricatives in Hong Kong English should be retained.

9.2.2 Prevalence of variation of Hong Kong English fricatives

Previous studies on Hong Kong English phonology have pointed out variation of the realisations of fricatives such as TH-fronting, TH-stopping, and voiced fricatives being substituted by voiceless fricatives. Nevertheless, many of the studies (except Hansen Edwards, 2019) did not provide frequency counts of the observations. This study attempts to answer how prevalent variation is in the phonology of Hong Kong English, in general. Table 9.23 shows the prevalence of variation per fricative phoneme and the distributions of the variants.

Table 9.23 Prevalence of variation per fricative in the dataset

Fricative	variation	variant(s)
/f/	0%	NA
/v/	94.3%	[f]: 93.6% [w]: 6.4%
/θ/	32.4%	[f]: 100%
/ð/	96.4%	[θ]: 53.7% [d]: 23.6% [f]: 22.7%
/s/	2.2%	[ʃ]: 100%
/z/	88.8%	[s]: 97.4% [ʃ]: 2.6%
/ʃ/	1.8%	[s]: 100%
/ʒ/	96.4%	[ʃ]: 49.5% [ʧ]: 46.3% [s]: 4.2%

As can be seen, for the voiced fricatives /v, ð, z, ʒ/, the percentages of variation observed in the dataset are very high. It is because the phonetic voiced fricatives rarely occurred in the data set. Nevertheless, instead of four phonemes, the present study proposed eight phonemes for the inventory of Hong Kong English fricatives. For the voiceless fricatives, no variation for /f/ is observed. The percentages of variation observed for /s/ and /ʃ/ are very low, meaning they are mostly realised as the canonical /s/ and /ʃ/.

The prevalence percentages help decide which variation is representative in Hong Kong English and is worth further investigation. For example, Bolton and Kwok (1990, p. 153) reported the list of observations of substitution, as shown below:

a. RP /θ/ is replaced by /f/, e.g. [fiŋ] *thinks*

b. /ð/ is replaced by /d/ in initial position, e.g. [deɪ], and by /v/ in final position, [wiv̥] *with*

c. /v/ is replaced by /w/, e.g. [d̥iwaɪ], *divide*

d. /ʃ/ is replaced by /s/, e.g. [iŋglis], *English*

In the present study, the variant [s] of /ʃ/ proposed by Bolton and Kwok (1990) only occurs 1.8% among the total number of /ʃ/. It can be inferred that this observation is not representative, at least based on the data of the current study. In the next section, the variants are discussed with respect to different linguistic factors.

9.2.3 Variation and linguistic factors

Variation of /v/

Regarding the variation of the voiced labiodental fricative /v/ (94.3%), the present study found that it is mostly realised as [f]. Nevertheless, syllable onset position slightly favoured [v] but the overall occurrences of [v] are still low. A small number of the variant [w] are also noted in the syllable onset position but stress is not a significant predictor. It can be postulated that the realisation as [w] may be idiosyncratic. 28 speakers pronounced /v/ as [w]. Some speakers produced such a variation consistently, meaning [w] is an allophone of /v/ in their phonological systems. Others produced a mixed [w] or [f], meaning they behave more like free variants. In general, the present study found that:

(i) [v] is observed although the occurrence is marginal

(ii) /v/ is mostly realised as the voiceless counterpart [f] in all environments

(iii) The realisation as [w] (in syllable onset position) does not occur frequently

The findings partially confirm the predictions stated in Chapter 5: the main variants of /v/ will be the voiceless [f] and the voiced labialised velar approximant [w] (Bolton and Kwok, 1990; Hung, 2000). Only the first part is confirmed. In addition, although it can be claimed that the main variant of /v/ is [f], given the high percentage of substitution of [f] for /v/, it can also be said that /v/ is primarily realised as [f], while the realisations as [v] and [w] are marginal.

Variation of /θ/

Regarding the variation of the voiceless dental fricative /θ/, the present study noted that the variation is around two-third (32.4%). That is to say, the majority (67.6%) of the production of /θ/ are [θ]. With respect to the variation, this study found that [f] is more likely to occur in syllable coda position and unstressed syllable respectively. In general, the present study found that:

(i) Around two-third of /θ/ are realised as [θ]

(ii) There is only variant [f], but no instances of [s]

(iii) The variant [f] is more likely to occur in syllable coda position than onset position

(iv) Preceding labial consonant is a not a significant predictor

The findings confirm the predictions stated in Chapter 5 that two-third of the realisations of /θ/ will be [θ]. This finding is different from what Hansen Edwards (2019) reported that 46% of the production are [θ] while the other 54% is variation. The finding of this study is more in line with the findings in the studies by Hung (2000) and Deterding et al. (2008), in which two-third of the target /θ/ were realised as [θ].

Nevertheless, the prediction: the majority of the variants of /θ/ will be [f] and only a small proportion of variants will be [s] cannot be confirmed. Hansen Edwards (2019) noted 5% (n = 84) of the realisation of /θ/ was [s] and found that English proficiency had an effect on TH variation that speakers with advanced level were more likely to produce the variant [s]. Nevertheless, in the present study, only 13 out 3329 (0.39%) observations of /θ/ are realised as [s]. Hence, this variant was discarded from modelling. The finding of this study is more in line with the findings in the studies by Bolton and Kwok (1990), Hung (2000), and Deterding et al. (2008).

The findings of the present study also cannot confirm the prediction: TH-fronting is more likely to occur when there is a preceding labial consonant in the same word. Hansen Edwards (2019) found that the presence of a preceding labial consonant in the same word was more likely to trigger the variation but no main effect for preceding labial consonants is found in the present study.

Variation of /ð/

Regarding the variation of the voiced dental fricative /ð/, the present study found that the variation was very high (96.4%). There are three variants, namely [θ], [d], and [f]. [θ] is more likely to occur in coda position and [d] is more likely to occur in syllable onset position, but they are not complementarily distributed. There was a significant amount of [θ] occurring in syllable onset position as well. There was no significant predictor found for the realisation of /ð/ as [f]. In general, the present study found that:

(i) [ð] is observed although the occurrence is marginal

(ii) The variant [d] is more likely to occur in syllable onset position

(iii) The variant [θ] is more likely to occur in syllable coda position but there are still a large amount of [θ] in syllable onset position

(iv) The variant [f] occurs in all environments but with relatively low frequency

The findings of the present study partially confirm the predictions stated in Chapter 5: the realisation of /ð/ as [d] (TH-stopping) is more likely to occur in syllable onset position and the variant [f] is more likely to occur in syllable coda position. Syllable position has no effect on the occurrences of variant [f] of /ð/. Overall, the occurrences of [f] are low. The findings tend to confirm what was proposed by Hung (2000) and Bolton and Kwok (1990) that [d] was pronounced in word-initial and intervocalic position. Nevertheless, the realisation of /ð/ as [θ] should not be ignored.

Variation of /z/

Regarding the voiced alveolar fricative /z/, the substitution rate is very high (99.8%) and most of the /z/ are realised as the voiceless [s]. In this case, it can be postulated that there is almost no variation. Only a small amount of /z/ are realised as [ʃ]. The realisation of /z/ as [ʃ] is more likely to occur when preceded by back rounded vowel /u/ and in unstressed syllables. Some speakers were more susceptible to the influence of /u/ when pronouncing /z/, and hence, produced a more anterior like fricative, which is the voiceless postalveolar fricative /ʃ/ in this case. In general, the present study found that:

(i) [z] is observed although the occurrence is marginal

(ii) /z/ is mostly substituted by [s] in all environments

(iii) A small number of the variant [ʃ] of /z/ is observed and the variant is more likely to occur when preceded by the back rounded vowel /u/ and in unstressed syllables.

The findings of the current study cannot confirm the predictions stated in Chapter 5 that /z/ is more likely to be pronounced as [ʃ] when followed by back rounded vowels. The following /u/ has no effect on the occurrences of [ʃ] in this study. Nevertheless, there is an effect for preceding /u/. This is not surprising since co-articulation effect can also come from preceding context, as demonstrated by Jongman et al. (2000). The oral cavity is enlarged or extended due to the lip rounding of /u/, and hence, the postalveolar fricative is more likely to be produced, as explained in Section 3.1.

Variation of /ʒ/

The variation of /ʒ/ is very high (96.4%). The present study found that the variant [ʃ] is more likely to occur in syllable onset position and the variant [ʧ] is more likely to occur in syllable coda position. The variant [s] is more likely to occur in syllable onset position although the overall occurrence was very low. In summary, the present study found that:

(i) [ʒ] is observed although the occurrence is marginal

(ii) The main variants of /ʒ/ are [ʃ] and [ʧ]

(iii) The variant [ʃ] is more likely to occur in syllable onset environment

(iv) The variant [ʧ] is more likely to occur in syllable coda environment

This finding is different from the prediction stated in Chapter 5 and what Hung (2000) and Setter et al. (2010) claimed: all /ʒ/ is replaced by the voiceless /ʃ/.

9.3 Phonological rules of Hong Kong English fricatives

The key findings and discussions can be summarised by formulating phonological rules. Nevertheless, the phonological rule, per se, may be misleading since none of the devised rules is absolute (in terms of probability). Therefore, each rule is labelled as 'high', 'mid', or 'low', which refers to the probability of occurrence of such a variation. The phonological rules of Hong Kong English fricatives are proposed in Table 9.24.

Table 9.24 Proposed phonological rules of Hong Kong English fricatives

Fricative	Phonological rule	Probability of occurrence
/v/	/v/ → [f] in all environments	high
	/v/ → [w] in syllable onset position	low
/θ/	/θ/ → [f] in syllable onset position	low
	/θ/ → [f] in syllable coda position	mid-low
/ð/	/ð/ → [θ] in onset position	mid-high
	/ð/ → [θ] in syllable coda position	high
	/ð/ → [d] in syllable onset position	high
	/ð/ → [d] in syllable coda position	low
	/ð/ → [f] in syllable onset position	low
	/ð/ → [f] in syllable coda position	low
/z/	/z/ → [s] in all environments	high
	/z/ → [ʃ] in all environments	low
/ʒ/	/ʒ/ → [ʃ] in syllable onset position	high
	/ʒ/ → [ʃ] in syllable coda position	low
	/ʒ/ → [ʧ] in syllable onset position	low
	/ʒ/ → [ʧ] in syllable coda position	high
	/ʒ/ → [s] in syllable onset position	mid-low
	/ʒ/ → [s] in syllable coda position	mid-low

Chapter 10

Application in automated speech recognition (ASR) system

10.1 Results of MAUS adaptation

The findings from the auditory analysis and the phonological rules in Chapter 9 were transformed into pronunciation rules with *a prior* probabilities using the method described in Section 6.7. They were then applied in WebMAUS together with the modified standard British English (GB) pronunciation rule set to automatically label the phones from the subset of story data. For simplification purposes, this model is referred to as GB-HKE hereafter. The baseline model is the GB MAUS model with default pronunciation rule set for British English. The phones estimated by the models are called hypotheses (HYPs). The labels from the phonetic transcription by the researcher are called references (REFs). A total of 1815 target fricatives were examined. The accuracy was computed by comparing the hypotheses and references (accurate HYPs / total no. of entries). Another common metrics to evaluate phone recognition performance is to measure the Phone Error Rate (PER), which is also the Levenshtein distance (no. of insertions + deletions + substitutions) between the decoded output and the reference, normalised by the length of the reference at the phonemic level but not word level (Kong et al., 2017). Nevertheless, this study concerns isolated phones but not phones in sequence, therefore, PER was not adopted. The overall performance of target phone recognition is shown in Table 10.1. The phone recognition accuracy of the baseline GB model is 61.5%. The phone recognition accuracy of the GB-HKE model is 86.0%. An averaged improvement of 24.5% is achieved when using the GB-HKE model.

Table 10.1 Overall accuracy of target phone recognition in MAUS

(N=1815)	Baseline GB	GB-HKE	Improvement
Accuracy	61.5%	86.0%	24.5%

The phone recognition performance per speaker is illustrated in Table 10.2. For the performance of the baseline GB model, it ranges from 49.3% to 76.4%. As for the performance of the GB-HKE model, it ranges from 84.1% to 88.7%. Despite the small sample size, the improvements of phone recognition accuracy suggest that this set of pronunciation rules of Hong Kong English fricatives are able to boost the performance to at least 84% per speaker regardless how the baseline model performed.

Table 10.2 Accuracy of phone recognition by speaker

Speaker	Baseline GB	GB-HKE	Imrovement
F01	273/457 (59.7%)	391/457 (85.6%)	25.9%
F02	275/452 (60.8%)	380/452 (84.1%)	23.3%
M01	225/456 (49.3%)	391/456 (85.8%)	36.5%
M02	344/450 (76.4%)	399/450 (88.7%)	12.3%

Table 10.3 displays the misrecognised phones by the baseline GB model and the GB-HKE model. For the baseline GB model, altogether, there are 698 instances of misrecognition, distributed among 23 unique types. As for the GB-HKE model, altogether, there are 254 instances of misrecognition, distributed among 27 unique types. Since there is a long tail in the distribution of types, only the five most frequent types of misrecognition are reported in the table.

Table 10.3 Error analysis sorted by frequency of occurrence (Top 5)

a. Baseline GB

REF	HYP	*n*	%
s	z	249	35.7
f	v	174	24.9
d	ð	170	24.4
f	θ	29	4.2
θ	ð	20	2.9

(N = 698)

b. GB-HKE

REF	HYP	*n*	%
d	ð	83	32.7
s	z	45	17.7
θ	f	44	17.3
f	v	20	7.9
ð	d	9	3.5

(N = 254)

As can be seen from the error analysis of the baseline GB model, there seem to be a tendency to over-recognise voiceless fricatives as the voiced counterparts such as for voiced/voiceless alveolar fricatives (35.7%), labiodental fricatives (24.9%) and dental fricatives (24.4%). The confusion of /f/ and /θ/ (4.2%) as well as of /θ/ and /ð/ (2.9%) is also an issue for automatic phone recognition, though not a major one. Regarding the error analysis of the GB-HKE model, the types of most frequent misrecognition are similar to that of the baseline GB model. The confusion of /d/ and /ð/ (32.7%) is a major problem of recognition, followed by the confusion of /s, z/ (17.7%). There is a tendency to over-recognise /θ/ as /f/ (17.3%).

10.2 Discussion on the application in an existing ASR system

The following discussions attempt to answer the research question:

(i) To what extent can the findings of this study be applied to an existing state-of-the-art automatic speech recognition (ASR) system and improve the phone recognition of Hong Kong English fricatives and their variants?

10.2.1 Pronunciation rules

One of the aims of the present study is to apply the findings of this study in an existing state-of-the-art automatic speech recognition (ASR) system. By comparing the phone recognition performances, the generalisability of the findings can be evaluated. In this study, the fricative production using pseudo-words with highly controlled phonetic contexts embedded in a carrier phrase was examined. The findings were then transformed into pronunciation rules and applied in WebMAUS to automatically label the story reading data, in which real words were used. The difference between the highly controlled pseudo-words and the real words in the story is that the real words in the story comprise more different vowel and consonant contexts as well as more complex syllable structures than the pseudo-words in the word list. In terms of probability, the conditional probabilities of phoneme occurrences are also different in the two sets of data. In the highly controlled word list data, a simplistic approach was adopted insofar that the conditional probability of a phoneme given different sequence combination was assumed to be the same since the English phonotactics, syntax, and semantics were constant. In the story reading data, the sequence combinations and their conditional probabilities are more diverse. For example, the story was written in past tense, and hence, it is more likely for the morpheme -ed (/d/ and/or /t/) to appear in the word-final position.

The standard British English (GB) model in MAUS was employed in this study. Since the GB language and acoustic model was trained on the AIX-MARSEC corpus (Auran et al., 2004), which contained 55,000 transcribed words of spoken (standard) British English, most common phoneme sequences and their probable pronunciation variants are covered. Recall that the pronunciation model is an acyclic directed graph, in which the nodes represent a single phone symbol, and the arcs represent the transitions from one symbol to the next symbol with a conditional probability. What the pronunciation rules essentially do is to replace and add more arcs to the graph. MAUS supports two types of pronunciation rules: i) rules with statistical information and ii) rules without statistical information. Rules without statistical information implies that the conditional probability is one, which is an absolute transition arc. This type of pronunciation rule is not applicable for the present study because results from the auditory analysis of the word list data suggest that the variants of fricative production are not complete substitution or replacement. For example, even though the findings indicate that the realisation of /ʒ/ as [ʧ] is more likely to occur in syllable coda position, it can also be realised as /ʒ/ (although marginally) and /ʃ/. In this case, the arcs to /ʒ/ cannot be removed. Instead, the arcs to /ʧ/ and /ʃ/ needs to be added. Therefore, rules with statistical information were adopted in the present study.

As mentioned in Section 6.7, a ‘greedy’ approach was adopted when generating the MAUS pronunciation rules, meaning phonetic environments which appeared in the story but not in the word list were also included and the same conditional probabilities were applied. It also means that the original conditional probabilities were overwritten, and the new probabilities were estimated using idealised conditions. This estimation based on the limited phonetic environments in the word list data cannot truly represent the conditional probabilities estimated from a large corpus data. Nevertheless, for experimental purposes, the probabilities estimated from the word list data were applied. The phone labelling results of the baseline

GB model and the GB-HKE model were compared. The overall improvement of 24.5% of the target phone recognition (from 61.5% to 86.0%) suggests a similar pattern of the realisation of fricatives in the story data.

Results of the error analysis indicate that the number of voiced/voiceless fricative confusion as well as the confusion of /ð, d/ and /θ, f/ has significantly dropped. Despite the drop in overall misrecognition, these confusions comprise 79.1% of the misrecognition. One way to improve the recognition performance is to adjust the conditional probabilities per rule. In this case, more annotated story reading data or, in general, a corpus of spoken Hong Kong English, are necessary. The remaining types of misrecognition (20.9%) are relatively miscellaneous. It is because there were instances of mispronunciation, self-correction, hesitation, and disfluency in the story data. Misrecognitions due to these reasons are difficult to improve. That is to say, the overall phone recognition accuracy of the GB-HKE model can be increased to around 88.9%, at most. In this case, the overall performance of 86.0% using the idealised probabilities from the word list data suggests that the GB-HKE model is almost optimal.

To conclude, generating pronunciation rules using results of highly-controlled pseudowords and applying the devised rules in an existing state-of-the-art ASR system can improve the overall phone recognition accuracy of non-spontaneous speech data. Such an application is particularly useful when the target variety (e.g. many new varieties of English including Hong Kong English and low-resource languages) is not available or supported in the existing ASR systems. Instead of training an independent language and acoustic model for the target variety, the existing language and acoustic model of a language can be adopted and optimised for the target variety. In fact, training a language and acoustic model requires a large amount of annotated data and is, by no means, a trivial task (Schiel, 1999; Schiel, 2015; Kisler et al., 2017; Auran et al., 2004). Adding pronunciation rules of the target variety to an existing language and acoustic model is a practical solution.

It should be noted that in MAUS, the pronunciation rules use SAMPA symbols, and each language has a different set of symbols. Each MAUS model only accepts the SAMPA symbols of that language. Using unstated SAMPA symbols is not permitted. Moreover, since each acoustic model and language model of a language variety is trained independently, it can be assumed that the acoustic feature vectors associated with same SAMPA symbol are different, depending on the language variety. For example, the acoustic vectors of /u/ in the standard American English (US) model are slightly different from the standard British English (GB) model, and from the German (DE) model. The application of the pronunciation rules in existing ASR systems means the associated acoustic vectors of the chosen model will be used. In the next Section, the application of acoustic features in the ASR systems is discussed.

10.2.2 Acoustic and phonological features

In an ASR system, Mel-frequency cepstral coefficients (MFCCs) are the common features to be extracted as the acoustic vectors. In this process, a series of window frame (usually 25 ms) is applied on the speech signals. Within this time frame, the signals or waveforms are assumed to be static (C. Zhang, 2017). Apart from MFCCs, time derivatives estimated using linear regression coefficients within the window, are also added to the feature vector (Young et al., 2002). The output is a sequence of feature vector. In MAUS, the standard feature set of 12 MFCCs + Energy + the first and second time derivative is computed per each time frame (Kisler et al., 2017). It can be assumed that other ASR systems and forced alignment tools (e.g. FAVE-align) which use the hidden Markov Toolkit (HTK) also adopt similar pre-processing of audio files (see Section 4.2).

What is of more interest for the present study is how a phoneme (or a phone symbol) is treated in the acoustic model. Each phoneme is actually modelled by a hidden Markov model (HMM). It is widely known that the acoustic properties of a phoneme change from beginning to end (see also the figures of window location in Jongman et al., 2000 and Maniwa et al., 2009), and that there is a coarticulation effect from the preceding and following sound. Therefore, the phonemes are usually divided into several states. In MAUS, three states are employed for consonants and short vowels and four states are employed for long vowels and diphthongs. Each state represents the probability distribution of feature vector in the beginning, middle, and end of the phone. The probability densities of each state are modelled separately (Senior et al., 2015). In addition, since the nature of speech acoustics is sequential and forward-in-time, the HMM is restricted to forward transitions only (Šilingas and Telksnys, 2004). The transition probability between states essentially denotes how long the current state has been occupied. Figure 10.1 is an illustration of a three-state HMM for a phoneme. As can be seen, it is not necessary to parse all three states. Sometimes the state can be passed if the phones are pronounced very fast (Šilingas and Telksnys, 2004).

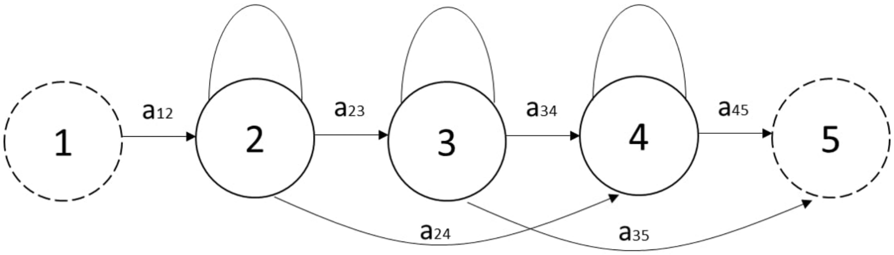

Figure 10.1 A three-state left-to-right HMM of a phoneme extracted from Šilingas and Telksnys (2004, p. 96)

In short, each state contains information of a feature bundle of MFCCs, energy and time derivatives, all derived from the speech signals. Altogether, they form an internal HMM of a phone. MFCCs are suitable data type for modelling since they contain adequate information which represents a phone segment (C. Zhang, 2017). Other features can also be used. For example, Senior et al. (2015) used 40-dimensional log mel filterbank features, from which MFCCs are derived. Nevertheless, MFCCs and log mel filterbank features are relatively opaque, meaning the acoustic characteristics of a phone segment cannot be explicitly explained.

In order to investigate the acoustic differences between phones, a multitude of properties can be compared. The present study examined 10 acoustic properties plus two DCT coefficients of Hong Kong English fricatives and their variants. Statistical models were also run to estimate the predictability of two most relevant groups of phonological features, namely place of articulation and voicing. The two phonological features facilitate the understanding of the acoustic characteristics of each phoneme since phoneticians generally acknowledge that a fricative with features [-voice][+alveolar] (or [+anterior][+strident]) is /s/. Results from the acoustic analysis of Hong Kong English fricatives and their variants show that several acoustic properties of different fricatives are significantly different from each other by place (83.4%) and voicing (96.1%). The classification models of place and voicing using MFCCs also demonstrate relatively good performance. It suggests that certain phonological features are useful when it comes to improving the phone classification performance.

The error analysis of the GB-HKE model demonstrates that the confusion mainly comes from /d, ð/, /s, z/, and /θ, f/. In terms of phonological features, the confusion of /d, ð/ can be interpreted as a confusion of continuant: /d/ is [-continuant], while /ð/ is [+continuant]. The confusion of /s, z/ can be interpreted as a confusion of voicing: /s/ is [-voice], whereas /z/ is [+voice]. The confusion of /θ, f/ can be interpreted as a confusion of place of articulation: /θ/ is [+dental] (or [+coronal][-labial]), while /f/ is [+labiodental] (or [-coronal][+labial]). As demonstrated in the present study, phonological features can be extracted from the input signals using neural networks. These phonological features, together with the MFCCs, can be used in the acoustic model. Unfortunately, this part cannot be manifested in the present study because it means each HMM of a phone needs to be re-trained with the new feature vectors, which is beyond the scope of this study. Nevertheless, previous studies have shown recognition improvement with similar implementation. For example, Deng and D. Sun (1994) built an HMM-based ASR system using articulatory features (e.g. lips, tongue blade, tongue dorsum, velum, and larynx) in order to model the context-dependent behaviours (i.e. coarticulation effect) in speech. J. Sun and Deng (2002) incorporated high-level linguistic constraints such as "word and phrase boundaries, morpheme, syllable, syllable constituent categories, and word stress" (p. 1086) in an HMM-based ASR system. Results showed that the overlapping-feature model performed better than the traditional context-dependent (triphone-based) acoustic model. Therefore, it can also be predicted that by applying neural networks of phonological feature extraction of Hong Kong English fricatives and their variants in an adapted ASR system, the phone recognition performance can also be improved.

Chapter 11
Conclusion

The present study systematically examined the acoustic characteristics and the phonology of Hong Kong English fricatives. As pointed out by Bolton and Kwok (1990), the phonological system of Hong Kong English is dynamic in a sense that it contains Hong Kong English phonological features as well as the British or American English features, and the distribution of features is different from speaker to speaker. If Hong Kong English phonology is to be conceptualised as a continuum, as proposed by Q. Zhang (2013) and Hung (2000), the ratio of endonormative features and exonormative features (of standard British or American English) is different when moving along the continuum. The idealised Hong Kong English phonology would demonstrate all localised features with high frequency, and vice versa for British or American English phonology. Nevertheless, neither the dynamic representation nor the continuum representation of Hong Kong English suggests how the phonological system of a typical Hong Kong English speaker is like. This study recruited 106 university students in order to construct a sample of mid-range Hong Kong English speakers. In line with previous studies, there is an assumption that a mid-range Hong Kong English speaker is a typical Hong Kong English speaker. The term 'mid-range' was first used by Bolton and Kwok (1990, p. 151). However, none of the studies of Hong Kong English phonology have attempted to answer what 'mid-range' actually denotes. Does it mean half of the phonological features are Hong Kong English and half of the phonological features are standard British/American features? Results of the current study reveal that the realisations of Hong Kong English fricatives in terms of exonormative and endonormative features are not at all fifty-fifty. Instead, most of the realisations of fricatives are considered as endonormative features. Examples include substituting voiceless fricatives are substituted for the voiced fricatives and [d] and [ʧ] to substitute for /ð/ and /ʒ/. That is to say, the 'mid-range' speakers in this study are more included to the idealised Hong Kong English phonology end in the continuum of Hong Kong English phonology rather than the idealised British/American English phonology end.

Regarding Schneider's (2007) Dynamic model, findings of the present study suggest that the pronunciations of fricatives are relatively consistent and predictable, although there are cases of free variation. This means that the realisations of Hong Kong English fricatives are "stabilised linguistically to a considerable extent" (Schneider, 2007, p. 51). A homogeneity of the realisation of Hong Kong English fricatives with respect to different linguistic factors (mainly syllable position) across speakers indicates that Hong Kong English is very much in the stage four–endonormative stabilisation in the Dynamic model.

A lot of emphasis has been put on the method and quantifying the observations into probabilities. In this regard, this study is different from previous studies on the phonology of Hong Kong English which suggested a number of phonological features but not much statistical information was provided. Some phonological features and linguistic factors suggested in previous studies cannot be attested in the present study. For example, replacement of /ʃ/ with [s] and realisation of /ʒ/ as [v] cannot be observed in the present study. The realisation of /v/ as [w] is also of low occurrence. The present study also found more variants for /ð/ and /ʒ/ than previous studies. The variation of fricative with respect to different explanatory variables was also modelled. Some results suggested in previous studies such as the variant [d] of /ð/ being more likely to occur in syllable onset position can be replicated in this study. Nevertheless, some findings such as preceding labial consonant facilitating the realisation of /θ/ as [f] cannot be replicated. Reasons why the results cannot be replicated are manifold. One reason is that the data of Hansen Edwards (2019) was a mixture of reading and spontaneous speech, while the present study used only word list data for the auditory analysis. It can be the case that the production of /θ/ is more susceptible to neighbouring sounds in spontaneous speech.

Acoustic analysis of the fricatives and their variants was also conducted. It was found that the patterns of some acoustic properties such as skewness and kurtosis were different from standard American English. The predictabilities of phonological features using the acoustic properties were also found to be different. It comes to the question whether there is a need to create an acoustic model for Hong Kong English. A classification model of phone symbols was built using the acoustic signals and neural network, and achieved overall good performance (80.6%). Nevertheless, the data preparation, which required a large amount of segmented phone data, was extremely time-consuming and labour-intensive. The present study demonstrated a practical solution by applying the weighted pronunciation rules in an existing ASR system. The target phone labelling accuracy increased significantly from 61.5% to 86.0%. In this case, the acoustic model of standard British English was adopted with the phonological rules specific to Hong Kong English fricatives, and no new acoustic model was required. In fact, the trained models and the weighted pronunciation rule set are one of the major contributions of the present study. They can be re-used for future research studies on Hong Kong English fricatives and can significantly reduce the time on data processing. The computational methods used in this study can also be applied to study other low-resource language varieties.

Last but not least, what the current study examined is just a tiny part of the phonological system of Hong Kong English fricatives, namely how fricatives are realised in a CVC syllable with respect to four vowels /i, e, u, a/ and how syllable position and stress pattern influence the realisation. How fricatives behave in consonant cluster and spontaneous speech, and if there are any social factors such as gender, English proficiency, and age which may affect the realisations of fricatives are not covered in this study. Nevertheless, they are all interesting research questions and can be explored in future studies.

References

Abadi, Martín et al. (2016). "Tensorflow: A system for large-scale machine learning". In: *12th USENIX symposium on operating systems design and implementation (OSDI'16)*, pp. 265–283.

Abdelatty Ali, Ahmed M, Jan Van der Spiegel, and Paul Mueller (2001). "Acoustic-phonetic features for the automatic classification of fricatives". In: *The Journal of the Acoustical Society of America* 109.5, pp. 2217–2235.

Abdelli-Beruh, Nassima B (2012). "Voicing and Devoicing Assimilation of French/s/and/z". In: *Journal of psycholinguistic research* 41.5, pp. 371–386.

Albawi, Saad, Tareq Abed Mohammed, and Saad Al-Zawi (2017). "Understanding of a convolutional neural network". In: *2017 International Conference on Engineering and Technology (ICET)*. Ieee, pp. 1–6.

Anjos, Ivo et al. (2020). "Detection of voicing and place of articulation of fricatives with deep learning in a virtual speech and language therapy tutor". In: *Proceedings of the Interspeech, Shanghai, China* 28.

Arora, Vipul, Aditi Lahiri, and Henning Reetz (2018). "Phonological feature-based speech recognition system for pronunciation training in non-native language learning". In: *The Journal of the Acoustical Society of America* 143.1, pp. 98–108.

Ashliman, DL (1996). *Folklore and mythology electronic texts.*

Auran, Cyril, Caroline Bouzon, and Daniel Hirst (2004). "The Aix-MARSEC project: an evolutive database of spoken British English". In: *Speech Prosody 2004, International Conference.*

Baayen, R Harald, Douglas J Davidson, and Douglas M Bates (2008). "Mixed-effects modeling with crossed random effects for subjects and items". In: *Journal of memory and language* 59.4, pp. 390–412.

Bates, Douglas et al. (2015). "Fitting Linear Mixed-Effects Models Using lme4". In: *Journal of Statistical Software* 67.1, pp. 1–48. DOI: 10.18637/jss.v067.i01.

Bayley, Robert (2002). “The quantitative paradigm”. In: *The handbook of language variation and change*. Ed. by J.K. Chambers and Natalie Schilling. Wiley Online Library, pp. 117–141.

Behrens, Susan and Sheila Blumstein (1988). “Acoustic characteristics of English voiceless fricatives: A descriptive analysis”. In: *Journal of Phonetics* 16.3, pp. 295–298.

Boersma, Paul (1993). “Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound”. In: *Proceedings of the institute of phonetic sciences*. Vol. 17. 1193. Amsterdam, pp. 97–110.

Boersma, Paul and David Weenink (2004). *introductory tutorial to PRAAT*.

— (2018). *PRAAT, a system for doing phonetics by computer (version 6.0.37)(computer program)*. URL: http://www.praat.org/.

Bolton, Kingsley (2002). *Hong Kong English: autonomy and creativity*. Vol. 1. Hong Kong University Press.

Bolton, Kingsley and Helen Kwok (1990). “The dynamics of the Hong Kong accent: Social identity and sociolinguistic description”. In: *Journal of Asian Pacific Communication* 1.1, pp. 147–172.

Bukmaier, Véronique and Jonathan Harrington (2016). “The articulatory and acoustic characteristics of Polish sibilants and their consequences for diachronic change”. In: *Journal of the International Phonetic Association* 46.3, pp. 311–329.

Buschfeld, Sarah and Alexander Kautzsch (2017). “Towards an integrated approach to postcolonial and non-postcolonial Englishes”. In: *World Englishes* 36.1, pp. 104–126.

Census and . Statistics Department (2016). *Snapshot of Hong Kong Population*. URL: https://www.bycensus2016.gov.hk/en/bc-snapshot.html (visited on 04/19/2021).

Chan, Alice Y.W. and David C.S. Li (2000). “English and Cantonese phonology in contrast: Explaining Cantonese ESL learners’ English pronunciation problems”. In: *Language Culture and Curriculum* 13.1, pp. 67–85.

Chan, Jim Y.H. (2016). “A multi-perspective investigation of attitudes towards English accents in Hong Kong: Implications for pronunciation teaching”. In: *Tesol Quarterly* 50.2, pp. 285–313.

— (2013). “Contextual variation and Hong Kong English”. In: *World Englishes* 32.1, pp. 54–74.

Chen, Wenda et al. (2010). “The development of a Singapore English call resource”. In: *Oriental COCOSDA, Nepal*.

Clark, Herbert H (1973). “The language-as-fixed-effect fallacy: A critique of language statistics in psychological research”. In: *Journal of verbal learning and verbal behavior* 12.4, pp. 335–359.

Deng, Li and Don Sun (1994). “A statistical approach to automatic speech recognition using the atomic speech units constructed from overlapping articulatory features”. In: *The Journal of the Acoustical Society of America* 95.5, pp. 2702–2719.

Deterding, David (2006). “The pronunciation of English by speakers from China”. In: *English World-Wide* 27.2, pp. 175–198.

— (2007). *Singapore English*. Edinburgh University Press.

Deterding, David, Jennie Wong, and Andy Kirkpatrick (2008). “The pronunciation of Hong Kong English”. In: *English World-Wide* 29.2, pp. 148–175.

Edge, Beverly (1991). “The production of word-final voiced obstruents in English by L1 speakers of Japanese and Cantonese”. In: *Studies in Second Language Acquisition*, pp. 377–393.

Ellbogen, Tania (2006). “Conventions for Segmentation”. In: *BAS Infrastrukturen zur Technischen Sprachverarbeitung (BITS)* Teilprojekt 8 (Doku 8/5e).Version 1.8.

Evers, Vincent, Henning Reetz, and Aditi Lahiri (1998). “Crosslinguistic acoustic categorization of sibilants independent of phonological status”. In: *Journal of phonetics* 26.4, pp. 345–370.

Fernandes, Joana et al. (2018). “Harmonic to noise ratio measurement-selection of window and length”. In: *Procedia computer science* 138, pp. 280–285.

Forrest, Karen et al. (1988). “Statistical analysis of word-initial voiceless obstruents: preliminary data”. In: *The Journal of the Acoustical Society of America* 84.1, pp. 115–123.

Gibbon, Dafydd, Roger Moore, and Richard Winski (1997). *Handbook of standards and resources for spoken language systems*. Walter de Gruyter.

Groves, Julie (2009). “Hong Kong English–Does it exist?” In: *HKBU Papers in Applied Language Studies* 13, pp. 54–79.

— (2011). “’Linguistic schizophrenia’in Hong Kong”. In: *English Today* 27.4, p. 33.

Guzik, Karita M and Jonathan Harrington (2007). “The quantification of place of articulation assimilation in electropalatographic data using the similarity index (SI)”. In: *Advances in Speech Language Pathology* 9.1, pp. 109–119.

Hansen Edwards, Jette (2015). “Hong Kong English: attitudes, identity, and use”. In: *Asian Englishes* 17.3, pp. 184–208.

Hansen Edwards, Jette (2016). “Sociolinguistic variation in Asian Englishes: The case of coronal stop deletion”. In: *English World-Wide* 37.2, pp. 138–167.

— (2018). *The politics of English in Hong Kong: attitudes, identity, and use.* Routledge.

— (2019). “TH variation in Hong Kong English”. In: *English Language and Linguistics* 23.2, pp. 439–468.

Harrington, Jonathan (2010). *Phonetic analysis of speech corpora.* John Wiley & Sons.

Harrington, Jonathan, Felicitas Kleber, et al. (2018). “Linking cognitive and social aspects of sound change using agent-based modeling”. In: *Topics in cognitive science* 10.4, pp. 707–728.

Harrington, Jonathan and Florian Schiel (2017). “/u/-fronting and agent-based modeling: The relationship between the origin and spread of sound change”. In: *Language* 93.2, pp. 414–445.

Hosom, John-Paul (2009). “Speaker-independent phoneme alignment using transition-dependent states”. In: *Speech Communication* 51.4, pp. 352–368.

Hui, Jonathan (2019). *Speech Recognition - GMM, HMM.* https://jonathan-hui.medium.com/speech-recognition-gmm-hmm-8bb5eff8b196. Accessed: 2021-06-17.

Hung, Tony TN (2000). “Towards a phonology of Hong Kong English”. In: *World Englishes* 19.3, pp. 337–356.

Jannedy, Stefanie and Melanie Weirich (2017). “Spectral moments vs discrete cosine transformation coefficients: Evaluation of acoustic measures distinguishing two merging German fricatives”. In: *The Journal of the Acoustical Society of America* 142.1, pp. 395–405.

Jelinek, Frederick (1997). *Statistical methods for speech recognition.* MIT press.

Jesus, Luis MT and Christine Shadle (2002). “A parametric study of the spectral characteristics of European Portuguese fricatives”. In: *Journal of Phonetics* 30.3, pp. 437–464.

Johnson, Keith (2011). *Acoustic and auditory phonetics.* John Wiley & Sons.

Jones, Daniel (2011). *Cambridge English pronouncing dictionary.* Cambridge University Press.

Jongman, Allard, Ratree Wayland, and Serena Wong (2000). “Acoustic characteristics of English fricatives”. In: *The Journal of the Acoustical Society of America* 108.3, pp. 1252–1263.

Kachru, Braj (1985). *Standards, codification and sociolinguistic realism: The English language in the outer circle.* Ed. by Kingsley Bolton and Braj Kachru. Vol. 3. London: Routledge, pp. 241–269.

Kisler, Thomas, Uwe Reichel, and Florian Schiel (2017). “Multilingual processing of speech via web services”. In: *Computer Speech & Language* 45, pp. 326–347.

Kong, Xiang, Jeung-Yoon Choi, and Stefanie Shattuck-Hufnagel (2017). “Evaluating automatic speech recognition systems in comparison with human perception results using distinctive feature measures”. In: *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pp. 5810–5814.

Kuznetsova, Alexandra, Per B. Brockhoff, and Rune H. B. Christensen (2017). “lmerTest Package: Tests in Linear Mixed Effects Models”. In: *Journal of Statistical Software* 82.13, pp. 1–26. DOI: 10.18637/jss.v082.i13.

Li, Fangfang, Jan Edwards, and Mary Beckman (2007). “Spectral measures for sibilant fricatives of English, Japanese, and Mandarin Chinese”. In: *Proceedings of the XVIth international congress of phonetic sciences*. Vol. 4, pp. 917–920.

Li, Fangfang, Benjamin Munson, et al. (2011). “Language specificity in the perception of voiceless sibilant fricatives in Japanese and English: Implications for cross-language differences in speech-sound development”. In: *The Journal of the Acoustical Society of America* 129.2, pp. 999–1011.

Lindstrom, Mary and Douglas Bates (1988). “Newton—Raphson and EM algorithms for linear mixed-effects models for repeated-measures data”. In: *Journal of the American Statistical Association* 83.404, pp. 1014–1022.

Luke, K.K. and J.C. Richards (1982). “English in Hong Kong: Functions and status”. In: *English World-Wide* 3, pp. 47–63.

Maniwa, Kazumi, Allard Jongman, and Travis Wade (2009). “Acoustic characteristics of clearly spoken English fricatives”. In: *The Journal of the Acoustical Society of America* 125.6, pp. 3962–3973.

Mitrović, Dalibor, Matthias Zeppelzauer, and Christian Breiteneder (2010). “Features for content-based audio retrieval”. In: *Advances in computers*. Vol. 78. Elsevier, pp. 71–150.

Nissen, Shawn L and Robert Allen Fox (2005). “Acoustic and spectral characteristics of young children’s fricative productions: A developmental perspective”. In: *The Journal of the Acoustical Society of America* 118.4, pp. 2570–2578.

Nittrouer, Susan, Michael Studdert-Kennedy, and Richard S McGowan (1989). “The emergence of phonetic segments: Evidence from the spectral structure of fricative-vowel syllables spoken by children and adults”. In: *Journal of Speech, Language, and Hearing Research* 32.1, pp. 120–132.

Paolillo, John (2002). *Analyzing linguistic variation: Statistical models and methods*. Center for the Study of Language and Inf.

Patgiri, Chayashree, Mousmita Sarma, and Kandarpa Kumar Sarma (2013). “Recurrent neural network based approach to recognize assamese fricatives using experimentally derived acoustic-phonetic features”. In: *2013 1st International Conference on Emerging Trends and Applications in Computer Science*. IEEE, pp. 33–37.

Pettarin, Alberto (2018). *Forced alignment tools*. https://github.com/pettarin/forced-alignment-tools/blob/master/README.md. Accessed: 2021-05-18.

Phung, Van Hiep and Eun Joo Rhee (2019). “A high-accuracy model average ensemble of convolutional neural networks for classification of cloud image patches on small datasets”. In: *Applied Sciences* 9.21, p. 4500.

R Core Team (2021). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria. URL: https://www.R-project.org/.

Rau, Victoria, Hui-Huan Ann Chang, and Elaine Tarone (2009). “Think or sink: Chinese learners’ acquisition of the English voiceless interdental fricative”. In: *Language Learning* 59.3, pp. 581–621.

Reichel, Uwe (2012). “PermA and Balloon: Tools for string alignment and text processing”. In: *Proc. Interspeech*.

Reichel, Uwe, Hartmut R Pfitzinger, and Horst-Udo Hain (2008). “English grapheme-to-phoneme conversion and evaluation”. In: *Speech and Language Technology* 11, pp. 159–166.

Roach, Peter (2009). *English phonetics and phonology paperback with audio CDs (2): A practical course*. Cambridge university press.

Schiel, Florian (1999). “Automatic phonetic transcription of non-prompted speech”. In: *Proceedings of the ICPhS*, pp. 607–610.

— (2015). “A statistical model for predicting pronunciation”. In: *Proceding of the International Conference on Phonetic Sciences*, p. 195.

— (2021). *WebMAUS General*. https://clarin.phonetik.uni-muenchen.de/BASWebServices/help. Accessed: 2021-05-21.

— (n.d.). *The Munich Automatic Segmentation System MAUS*. https://www.bas.uni-muenchen.de/Bas/BasMAUS.html.

Schiel, Florian et al. (1998). “The Partitur format at BAS”. In: *Proceedings of the First International Conference on Language Resources and Evaluation*.

Schneider, Edgar (2003). “The dynamics of New Englishes: From identity construction to dialect birth”. In: *Language* 79.2, pp. 233–281.

— (2007). *Postcolonial English: Varieties around the world*. Cambridge University Press.

Senior, Andrew, Haşim Sak, and Izhak Shafran (2015). "Context dependent phone models for LSTM RNN acoustic modelling". In: *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pp. 4585–4589.

Setter, Jane, Cathy S.P. Wong, and Brian H.S. Chan (2010). *Hong Kong English*. Edinburgh University Press.

Sewell, Andrew and Jason Chan (2010). "Patterns of variation in the consonantal phonology of Hong Kong English". In: *English World-Wide* 31.2, pp. 138–161.

Shadle, Christine (1985). "The acoustics of fricative consonants". In.

Šilingas, Darius and Laimutis Telksnys (2004). "Specifics of hidden Markov model modifications for large vocabulary continuous speech recognition". In: *Informatica* 15.1, pp. 93–110.

Stevens, Kenneth N (2000). *Acoustic phonetics*. Vol. 30. MIT press.

Stuart-Smith, Jane (2020). "Changing perspectives on /s/ and gender over time in Glasgow". In: *Linguistics Vanguard* 6.s1, pp. 1–13.

Sun, Jiping and Li Deng (2002). "An overlapping-feature-based phonological model incorporating linguistic constraints: Applications to speech recognition". In: *The Journal of the Acoustical Society of America* 111.2, pp. 1086–1101.

Sussman, Harvey M, Kathryn A Hoemeke, and Farhan S Ahmed (1993). "A cross-linguistic investigation of locus equations as a phonetic descriptor for place of articulation". In: *The Journal of the Acoustical Society of America* 94.3, pp. 1256–1268.

Sussman, Harvey M, Helen A McCaffrey, and Sandra A Matthews (1991). "An investigation of locus equations as a source of relational invariance for stop place categorization". In: *The Journal of the Acoustical Society of America* 90.3, pp. 1309–1325.

Viterbi, Andrew (1967). "Error bounds for convolutional codes and an asymptotically optimum decoding algorithm". In: *IEEE transactions on Information Theory* 13.2, pp. 260–269.

Watson, Catherine I and Jonathan Harrington (1999). "Acoustic evidence for dynamic formant trajectories in Australian English vowels". In: *The Journal of the acoustical society of America* 106.1, pp. 458–468.

Wickham, Hadley (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. ISBN: 978-3-319-24277-4. URL: https://ggplot2.tidyverse.org.

Winkelmann, Raphael et al. (2020). *emuR: Main Package of the EMU Speech Database Management System*. R package version 2.1.1.

Young, Steve et al. (2002). “The HTK book”. In: *Cambridge university engineering department* 3.175, p. 12.

Yuan, Jiahong, Wei Lai, et al. (2018). “Using forced alignment for phonetics research”. In: *Chinese Language Resources and Processing: Text, Speech and Language Technology. Springer.*

Yuan, Jiahong, Neville Ryant, et al. (2013). “Automatic phonetic segmentation using boundary models.” In: *Interspeech*, pp. 2306–2310.

Zhang, Chao (July 2017). “Joint Training Methods for Tandem and Hybrid Speech Recognition Systems using Deep Neural Networks”. PhD thesis.

Zhang, Qi (2013). “The attitudes of Hong Kong students towards Hong Kong English and Mandarin-accented English”. In: *English Today* 29.2, p. 9.

Zhao, Sherry Yi (2007). “The stop-like modification of /ð/: a case study in the analysis and handling of speech variation”. PhD thesis. Massachusetts Institute of Technology.

— (2010). “Stop-like modification of the dental fricative /ð/: An acoustic analysis”. In: *The Journal of the Acoustical Society of America* 128.4, pp. 2009–2020.

Appendix

Appendix A

Word List

Say bevdeð again.
Say ðetsheb again.
Say θiksib again.
Say wapʒad again.
Say fagzat again.
Say vapθak again.
Say puʒbuf again.
Say baftaθ again.
Say ðutshub again.
Say tuθpush again.
Say deðwep again.
Say teθpesh again.
Say zatfag again.
Say bivdið again.
Say kizbiv again.
Say piʒbif again.
Say peshkez again.
Say sabgas again.
Say diðwip again.
Say vupθuk again.
Say ʒudvup again.
Say shibðit again.
Say bavdað again.
Say pashkaz again.
Say wepʒed again.
Say zutfug again.
Say pushkuz again.
Say buftuθ again.
Say sebges again.
Say zitfig again.
Say fegzet again.
Say biftiθ again.
Say shubðut again.
Say fugzut again.
Say wipʒid again.
Say shabðat again.
Say θaksab again.
Say vipθik again.
Say peʒbef again.
Say ʒadvap again.
Say θekseb again.
Say ʒidvip again.
Say ʒedvep again.
Say kezbev again.
Say sibgis again.
Say guspuʒagain.
Say kazbav again.
Say shebðet again.
Say θuksub again.
Say buvduð again.
Say duðwup again.
Say gaspaʒagain.
Say wupʒud again.
Say paʒbaf again.
Say zetfeg again.
Say befteθ again.
Say vepθek again.
Say tiθpish again.
Say daðwap again.
Say ðatshab again.
Say gespeʒagain.
Say kuzbuv again.
Say pishkiz again.
Say subgus again.
Say taθpash again.
Say gispiʒagain.
Say figzit again.
Say ðitshib again.

Appendix B
Story

In this task, you will tell a story using the pictures given. The pictures were taken from a famous Disney movie called 'Snow White and the Seven Dwarfs'. The genre is fairy tale and the authors are Brothers Grimm from Germany. The passage below is an example. You can also devise or develop your version and add your own view. Let's read the following paragraphs first.

Once upon a time during the winter season, there was a queen who wished to have a child whose skin was as white as snow, with hair as black as the wood of the window frame, and lips as red as rouge. Soon after that in spring, she gave birth to a child as mentioned and the princess was therefore named Snow White. Three days later, the queen died because of fever and poor health condition.

On Snow White's seventeenth birthday, the king re-married to an attractive but wicked woman. The evil queen had a magical looking glass which only told the truth. She asked it who the most beautiful lady on earth was, and she assumed it was herself. However, its response was Snow White. The queen felt threatened and asked the huntsman to kill Snow White.

The huntsman took Snow White to the south, through the zigzag footpath, further down to the black forest. They crossed the bridge and arrived in front of a cottage in a small village. Then the huntsman was about to shoot Snow White. She was frightened and shouted for help.

The owners of the cottage, who were seven native dwarfs, were having lunch inside. They heard the noise and the screaming voice and saw the situation. They rushed outside and attacked the huntsman by throwing food like fish, shrimps, beef, cheese, fresh vegetables, strawberry, mushroom, spoons, forks, knife, and rubbish at him. The huntsman fled from the crazy assault and committed suicide as he realised the plan was sabotaged and screwed up, and the evil queen would send him to death anyways.

Thanks to the Seven Dwarfs, Snow White passed the crisis and her life was saved. To ensure her safety, they invited her to stay and live with them. During daytime, the Seven Dwarfs went out to work as usual, but they advised Snow White to make sure the door was locked and not to open for strangers easily.

Five months later, the queen asked the looking glass again, and to her surprise, its answer was still the same. She was shocked to find out that Snow White survived. She decided to kill her again with a more refined design: a poisonous apple.

She soaked the apple in a soup made with blood of presumably one thousand zombies and pythons in South East Asia, and mixed with vinegar, and a special spell enchanted by a wizard from voodoo. She disguised herself by dressing like a poor old lady zipped in black clothing and venture to fetch Snow White.

Snow White was at leisure when she heard someone knocking but she always remembered the advice and refused to open the door. The evil queen said, 'I am just a poor old lady who wish to share a very juicy apple with you and spread love. Let me give it to you through the window.'

The apple looked so juicy that Snow White thought it was worth trying. She took a bite of it in her mouth, then she zoned out.

The Seven Dwarfs thought Snow White was dead as she neither moved nor breathed. They put her in a glass coffin and held the funeral on top of the mountain. It happened that a prince of prestige from a distant country walked past. He instantly fell in love with Snow White and decided to give her a kiss out of sympathy.

To everyone's surprise, the kiss broke the spell. Snow White was alive again and they lived happily ever after.

Plagiatserklärung

[Bisher am Englishen Seminar verwendeter Text:]

Laut eines Vorstandbeschlusses des Englischen Seminars muss als letzte Seite der Arbeit folgende Erklärung abgegeben und unterschrieben werden:

Hiermit versichere ich, dass die vorliegende Arbeit über

…...

selbständig verfasst worden ist,
dass keine anderen Quellen und Hilfsmittel als die angegebenen benutzt worden sind und dass die Stellen der Arbeit, die anderen Werken- auch elektronischen Medien- dem Wortlaut oder Sinn nach entnommen wurden, auf jeden Fall Angabe der Quelle als Entlehnung kenntlich gemacht worden sind.

…..

(Unterschrift)