Intelligent Automation & Soft Computing
DOI:10.32604/iasc.2022.022180
Article

Tech Science Press

An Analysis of Perceptual Confusions on Logatome Utterances for Similar Language

Nur-Hana Samsudin[1,*] and Mark Lee[2]

[1]School of Computer Sciences, Universiti Sains Malaysia, 11800, Penang, Malaysia
[2]School of Computer Science, University of Birmingham, B15 2TT, Birmingham, UK
*Corresponding Author: Nur-Hana Samsudin. Email: nurhana.samsudin@usm.my
Received: 30 July 2021; Accepted: 31 August 2021

Abstract: In a polyglot speech synthesis, it is possible to use one language resource for another language. However, if the adaptation is not implemented carefully, the foreignness of the sound will be too noticeable for the listeners. This paper presents the analysis of respondents' acceptance of a series of listening tests. The research goal was to find out in the absence of phonemes of a particular language, would it be possible for the phonemes to be replaced with another language's phonemes. This will be especially beneficial for under-resourced language either in the case for 1) the language has not yet well researched into or 2) the language has not well documented in the required media. Preliminary studies were conducted to construct phoneme confusion matrices. The confusion study was observed based on the consonants' position in syllable structure: onset and coda. These studies were then compared to similar studies to find possible overlap among them. Then, based on the outcome, two perceptual tests have been conducted to observe the applicability of phoneme substitutions. The first test was to observe the effect of phonemes substitution during the intelligibility test for individual words. The second test was to evaluate whether context influenced perception based on whether respondents noticed phoneme substitution on a word in a series of words. From these experiments, it can be concluded that it is possible to do phoneme substitution but with a certain condition. From significance testing, it was found that phoneme substitution may not be suitable to be implemented for onset position but can be applied for coda position provided the context is available.

Keywords: Language similarity; phoneme analysis; polyglot speech; under-resourced language; natural language processing; speech analytics

1 Introduction

Creating a text-to-speech system (TTS) is no longer a difficulty for well-documented and well resource language with various speech analytic tools and abundance of studies in the last few decades. However, if the target language is not well documented and having insufficient digital resource, building a TTS would require supplementary resource from other languages. One of the limitations in producing speech

synthesisers from another language is the lack of phonemes available to facilitate the target language. For instance, there is the sound /β/ in Spanish but mainly perceived and produced like a/b/ by second or third language speakers. There is also the sound /r/ in French but there is also/r/ in French in which, only either one occur in most Asian languages. Therefore, even when one language's voice recordings can be used to produce another language's synthesiser; given that the foreignness of the synthesised speech is acceptable; the missing phonemes cannot be easily substituted with other phonemes for which (let's say) the manner or the place of articulation is near. This is one of the issues faced when some phonemes do not exist in the resource language.

Making a new recording for new (or non-existing) speech resources is not straightforward. A lot of things need to be considered including the repetitive recording process and the text preparation. These processes are also influenced by other parameters, for example the speaker's condition, device condition and background noises. These parameters, combined with a thorough procedure can create a lot of restrictions in creating acceptable data resources. Therefore, reusing an existing resource might be a preferable choice before a very elaborate recording and processing can be carried out.

The focus of this paper is on the analysis of phoneme confusion on logatomes utterance or also means nonsense utterances. The first section will give an overview of phoneme confusion, followed by the study on phoneme confusion done by multiple languages by different researchers. Then the study on phoneme confusion by multilingual listeners and an experiment on phoneme substitution will be presented.

2 Previous Studies on Phoneme Confusion

This paper attempts to identify the phoneme which can be perceived as another at the onset and coda position. This paper only consider single consonants and will not include complex onset or code which can be form by multiple consonants in the respective positions. This is to ensure the focus of the phones' perception is not influenced by multiple phones. Several studies have been conducted on human and machine perceived phonemes, among them are Miller et al. [1], Fant et al. [2], Lovitt et al. [3], Pinto et al. [4], Meyer et al. [5], Cutler et al. [6] and Christiansen et al. [7].

2.1 Study by Miller and Nicely (1955)

Miller et al. [1] used 16 consonants for identifying phoneme confusion by constructing logatomes utterances with a CV syllable constructed where the vowel is always /a:/. Miller et al. [1] devised confusion analysis to understand how humans confuse phonemes. What had been found in Miller et al. [1] was further studied by Fant et al. [2], Cutler et al. [6], Meyer et al. [5] and Pinto et al. [4]. Miller et al. [1] uses multiple speech frequencies with added noise to evaluate the effect of frequencies on perceived sounds. Five female subjects were used both as the speaker and as listeners. Four subjects are from the United States and one from Canada. The first language of the listeners however were not disclosed.

2.2 Study by Fant et al. (1966)

Fant et al. [2] used similar syllable structure to Miller et al. [1] where an English utterance test was constructed using 22 possible consonant phonemes at the initial position. Fant et al. [2] also conducted similar study for Swedish, where 17 possible initial single consonants were used. In their report, Fant et al. [2] also used logatome utterance, however they are abled to classified the language used into their respective languages. In both tests, they use one speaker and one listener to ensure consistent feedback. The listener was a bilingual with equal command of English and Swedish since childhood. The listener was given 10 randomised word lists for each language for each phoneme.

The confusion matric study of Fant et al. [2] listed the confusions that happened during the listening test in two conditions. In one, Fant et al. [2] listed the confusions that happened when listener was asked to hear a

recording which underwent low-pass filtering at 2000 Hz with a high quality filter. In the second, Fant et al. [2] presented the confusions that happened when white noise was added to 13 signal-to-noise ratio (SNR) sounds. The sounds were played over high quality loud speakers to the listeners. Due to the effect of low pass filtering on dentals and fricatives which resulted in those not being recognised at all, the results of added white noise were used as comparison. 13 dB noise is below the average speech level and therefore the effect of the noise was less drastic than the filtering [2].

2.3 Study by Cutler et al. (2004)

Cutler et al. [6] on the other hand expanded the study originated by Miller et al. [1] by using 24 consonants over 15 vowels used in English among 16 native listeners, and 16 non-native (Dutch) listeners. Cutler et al. [6] conducted a study using CV and VC structures and compared the confusion between American-English and Dutch speakers. The main focus of the study was to provide a new data set of phonetic identifications given a different level of noise (calculated by SNR) by native and non-native listeners. Cutler et al. [6] obtained 645 logatome syllables representing each of the different phoneme combinations. The noises were added from conversational speech which was also pre-recorded in a quiet room. Conversational speech was later added as a background noise to the recording. The recordings were mixed and added so that each logatome would have three different SNRs (0 dB, 8 dB and 16 dB). The results of the confusion matrices by Cutler et al. [6] has shown that the non-native listener performed below native phoneme-identification levels. However, Cutler et al. [6] also concluded that the non-native listeners appeared to remain constant (in producing the confusion phoneme) across SNRs with in the tested range as compared to native speakers.

2.4 Study by Meyer et al. (2007)

Meyer et al. [5] presented the comparison of human and machine phoneme recognition. In the human speech recognition test, Meyer et al. [5] used two kinds of signal. One was using noisy speech samples in which the sound to be evaluated was re-synthesised using MFCC. Another one used the original signal with added noise that was used to evaluate the loss of information caused by the process of re-synthesis. In their study, Meyer et al. [5] used CVC or VCV structures and, like Miller et al. [1], used nonsense utterances. For human speech recognition, five normal hearing listeners were requested to identify the utterance using the two types of signals given. 150 utterances were given to be evaluated. According to Meyer et al. [5] the choice of SNR when involving noise addition was based on presentation of only a few test lists to one human listener and proved to be reasonable for other test subjects as well. This was close to the SNR selected by Fant et al. [2] who chose to include an SNR of 13 dB.

2.5 Study by Lovitt et al. (2007)

Pinto et al. [4] in a different approach tried to identify where the causes of confusion started or happened in an automatic speech recognition system. Pinto et al. [4] extended the experiments in Lovitt et al. [3] which used only the CV structure by adding the VCV structure into the experiments. However, instead of human identification, Pinto et al. [4] used human mispronunciation, speech features confusion and phoneme recogniser confusion.

Pinto et al. [4] studied the confusion that occurred across three stages. Each confusion was categorised as the following: pronunciation confusion, frame confusion and phoneme confusion respectively. These were the three of the five stages in phoneme recognition. Pronunciation confusion refers to the mispronounced word. Frame confusion is the probability of error that the extracted features of the corresponding phonemes were not done correctly. Finally, the phoneme confusion is the mistaken identification by the phoneme recogniser itself. The purpose of the study by Pinto et al. [4] was to identify the confusion patterns to improve the performance of a recogniser by eliminating problematic phoneme distinctions.

Pinto et al. [4] wanted the phoneme recognition to be re-analysed into a smaller subset of phonemes which could be considered as common confusion patterns so that the system should be able to provide the supposed result and not to treat these selected phoneme group confusions as errors in phoneme identification. Pinto et al. [4] also stated that the confusion (from the phoneme recogniser) may have lost its voicing and place of articulation features which resulted in the misidentification of phoneme. This also conform with the direction of this paper whereby, when the voicing and the place of articulation information were lost, the phoneme can still be identified based on context.

2.6 Other Comparable Studies

Karanasou et al. [8] studied on keyword spotting aimed at detecting speech segments that contained a given query with in large amounts of audio data. One of the challenges of keyword spotting is how to handle recognition errors and out-of-vocabulary terms. This work proposed the use of discriminative training to construct a phoneme confusion model, which expanded the phonemic index of the keyword spotting system by adding phonemic variation to handle the recognition and out-of-vocabulary terms issues.

Žgank et al. [9], addressed the topic of defining phonetic broad classes needed during acoustic modeling for speech recognition during decision tree based clustering. A new data-driven method is proposed for the generation of phonetic broad classes based on a phoneme confusion matrix. The similarity measure is defined using the number of confusions between the master phoneme and all other phonemes included in the set. Žgank et al. [9] method, phonemes were classified into particular classes according to their similarity, determined by phoneme confusion matrix. The advantage of the defined method is that no expert knowledge is needed, which is often unavailable and can introduce subjective influence thus making it an advantage for multilingual speech recognition approach. They found that the proposed data-driven method improved speech recognition results when compared to the method based on expert knowledge.

Leijon et al. [10] presented a parametric Bayesian approach to the statistical analysis of phoneme confusion matrices measured for groups of individual listeners. Their study was to find out whether a new signal-processing system provided better phoneme recognition than a state-of-the-art reference system. Closely imitating Miller et al. [1], each participant might listen to a speech test material using different speech-coding algorithms in a cochlear implant system, or hearing aids adjusted using different fitting principles. Their study provide the insight of confusion studies in different speech recogniser environment may not directly tallied with the focus of this paper but one can find the comparison of human and machine confusion studies in this literature.

The analysis by Shi et al. [11] was not on speech recognition however it is relatable where they were researching the effect of first language on the ability to comprehend speech utterance. The performance evaluation was scored on words and phonemes: word-initial consonants (onset), vowels, and word-final consonants (coda). However, their confusion analysis were clustered based on the most dominant language of the listeners. Three type of listeners were English monolingual, English dominant (proficient in more than one language) and Russian dominant (non-native listeners). In their conclusion, both first-language phonology and second-language learning history affect word and phoneme recognition. They were hoping that their findings may help clinicians differentiate word recognition errors due to language background from hearing pathologies.

3 Preliminary Study on Phoneme Confusion

Preliminary studies were conducted to construct phoneme confusion matrices. For our phoneme confusion study, 255 sounds which consisted of CV, VC and CVC syllable structures are to be evaluated. Although these are logatome utterances, the purpose of our overall research is on finding if phoneme substitution is possible to be applied in a TTS system using other language's resource. For this

preliminary observation, Malay TTS is selected as a focal language for the synthesiser. The number of phoneme available in a particular language is dependent on different resources. There were 39 identified phonemes and two unidentified ones (due to the adaptive features of the phonemes when used in Malay) based on Ranaivo et al. [12], 34 phonemes were based on MBROLA-Group [13] and there were 38 based on Li et al. [14]. The phoneme lists were different due to the acceptance of declaring the borrowed phonemes from other languages. For example, the Malay phonemes listed by MBROLA did not include /v/ as a phoneme even though there are Malay words using this phoneme. This is because the loan words with such phonemes usually undergo transformation. For example, violin is known as 'biola' and is a loan word from Portuguese, 'viola'; goddess, is known as 'dewi' and is a loan word from Sanskrit, 'devi'; fasting, is known as 'puasa' pronounced as /puwasə/ and was a loan word from Sanskrit, 'upavasa'. However, for loan words from English, there were two categories, unplanned adaptation and planned adaptation [15,16]. For planned adaptation, in occurrences of /v/, slight changes took place; governor is 'gabenor' and private is 'prebet'. For unplanned adaptation, the words did not undergo transformation when there was a /v/. For example, television is 'televisyen', activity is 'aktiviti' and university is 'universiti'.

Additionally, in pronunciation, there was a slight variation which was also not listed. For example, Clyness and Deterding [17] stated that there is only one alveolar trill, "r", in Malay. However, during their observation of a speaker's recording, two "r"s were used: /r/ and /ɾ/. According to Clyness et al. [17], the speaker used the formal style. The speaker may have phonological influences from Standard Malay and English. Because there were no specific rules as to when a certain sound should be tap or trill, or when some audible release and reduction of a phoneme was supposed to take place, this paper only focus on one /r/ instead.

For this preliminary study, each consonant was paired to a vowel. Three vowels were used for this study to get a better description of human perception. Therefore there were three instances for each generated consonant. Each consonant was paired to one closed vowel, one mid vowel and one open vowel. Each sound was not supposed to be meaningful in Malay[1]. Malay vowels are not as easily confuse as some languages. For example in Bhatt et al. [18], the confusion happened in Hindi is mainly due to vowels and special analysis on confusion matric needed to be done for their vowels only.

This preliminary study record Malay confusion matrix based on 17 respondents. All respondents were multilingual and that include proficiency in Malay. This number of respondents was very close to the study conducted by Cutler et al. [6]. This number was different when compared to Miller et al. [1] and Meyer et al. [5] that both uses 5 respondents. Therefore the approach of creating the confusion matrices was more closely similar to the one proposed by Cutler et al. [6]'s than Miller et al. [1]'s. Pinto et al. [4] on contrast, used a phoneme recogniser to identify the confusion. According to Pinto et al. [4], the phoneme recogniser was making similar errors to a human speaker's made in speech identification.

All respondents were encouraged to take as long as they wished to answer, and allowing submission part-by-part so they were not stressed during listening. However, they were requested to use the same equipment to ensure the consistency of the given feedback. Contrary to the studies conducted by Miller et al. [1], Cutler et al. [6], Meyer et al. [5] and Pinto et al. [4] that used human speech and human speech recording, this study used synthesised speech. For the observation on phoneme confusion matrix, the generated sounds excluded the following consonants: /x/, /ʕ/ and /ʔ/ due to limited occurrences in the Malay training data itself.

3.1 Phoneme Confusion Matrix for Consonants in Syllable CV

This paper address the confusion happened for syllable CV, VC and CVC. The first was the study on phoneme confusion for consonants positioned at the beginning of the syllable CV (onset consonant).

[1]Five words coincidentally exist in Malay: kek, gam, tak, Mac and di.

The confusion matrix for CV is as presented in Tab. 1. As mentioned previously, each phoneme was paired with three vowels in different occurrences. The vowels used were: /a/, /e/ or /ə/ and /i/.

Table 1: Phonemes confusion for onset consonants for syllable structure: CV

Speech Synthesiser's Phones Production	Observed Phoneme Identified by Listeners																					
	b	tʃ	d	f	g	h	dʒ	k	l	m	n	ŋ	ɲ	p	r	s	ʃ	t	v	w	j	z
b	32		1	1						4				1	1				10	4		
tʃ		47															7					
d	1		46		1		1		1									1	1		1	1
f				50										1		1	1	1				
g		1	6		31		7	3											1			5
h			1	3	3	24	1	10	1					4				1			6	
dʒ		1					53															
k				1		4		34	12		3											
l									47		3				4							
m	2								2	38										11	1	
n									2	2	30	12								8		
ŋ					2						7	34	5								6	
ɲ									3	2		13	18							10	8	
p	12			21				1		1				16	1				1	1		
r	2		1	2										1	36					10		2
s		1														50	3					
ʃ		6														2	46					
t		4	6				9	2								5		26	1			1
v	1															1			39	10	3	
w									2	1			2		2				1	41	5	
j			2			1			10				2		1					2	36	
z			3		2		13									2						34

As shown in Tab. 1, /p/ was highly confused with /b/ and /f/ among Malay listeners. In fact, it has a greater impression of being an /f/ than /p/ itself. When compared against [1] however, the phoneme /f/ was only minimally confused as /p/ compared to: /k/, /t/, and /T/. Compared to Cutler et al. [6]'s experiment for the consonant in CV structures, /p/ was also confused among listeners with /b/ and /f/. But American English listeners also confused the /p/ to be /h/ which was also mistakenly identified as higher than the listeners' labelling of the /p/ as /p/ itself. As for Dutch listeners, /p/ was frequently heard as itself. It was also mostly confused with /h/, /b/, /f/ and /k/. Based on Meyer et al. [5], /p/ was highly confused with /k/, /b/, and /v/, and according to Pinto et al. [4], /p/ was confused with /b/, /t/, /k/ and /f/.

The detailed comparison across different approaches and studies are presented in Tab. 2. The main focus of this comparison was to see the phonemes which were noticeably identified as another. The different degrees of misidentification were shown in colours. The red showed that the produced phonemes were confused by being the perceived phonemes more than the correctly identified phonemes except for Pinto et al. [4]. The blue colour represented a different frequency number across the study.

Blue in the Malay study means the confusion was misidentified and happened not due to one person's miss-perception. For Miller et al. [1], the blue colour phonemes mean they were misidentified by more than ten times and greens showed that the misidentification happened ten or less but greater than or equal to four. This is due to the numbers involved in Miller et al. [1]'s study being very high and when misidentification of less than four happened, it is believed that it was caused by isolated mistakes. For Cutler et al. [6], blue referred to the number less than the number identified by the phoneme itself but higher than five, while green was for four or five frequencies of response only for the same reason as Miller et al. [1]. The same applied to Meyer et al. [5].

Pinto et al. [4], in contrast, did not use frequency of response. Therefore, the red in Pinto et al. [4] referred to the phoneme being misidentified throughout the three mentioned stages: human pronunciation confusion, frame speech features confusion and phoneme confusion. Blue indicates that the confusion happened in any of the two stages while green indicates that the misidentification happened only in one stage.

Table 2: Comparison of phonemes confusion across different observations for syllable CV

Phoneme	Malay	Miller and Nicely (1955)	Cutler et al. (2004)	Meyer et al. (2007)	Lovitt et al. (2007)
p	f, b	k, t, θ, f	h, f, b, k, θ, t h, b, k, f, t, v	b, k, v, g, t, f	t, k, f, b
b	v, m, w	v, ð, f, θ, d	h, m, ð, θ, f, v h, f, m, p, k, w, l, v	v, g, p	v, p, θ, d
t	dʒ, d, s, tʃ	p, k	h, p, k, θ, f p, k, h, θ, f, b	d, b	d, p, k, r, q, tʃ, s
d	(none)	g, ʒ, z, ð	n, ð, b, θ, j, l ð, n, l, b, h, j, θ, m	g, b	t, θ, g, dʒ, ɾ
tʃ	ʃ	(not tested)	t t, dʒ	(none)	ʃ, dʒ, t, s
dʒ	(none)	(not tested)	ð, tʃ, d tʃ, ð, d, j, p	(not tested)	ʒ, tʃ, z, j, d, t
k	l, h	(not tested)	h, t, p p, h, t	g, v	t, p, g
g	dʒ, d, z	d, ʒ, z, ð	j, h, n j, b, h, k	k, v	k, d, t
m	w	n	v, l, n n, b, r, l	l, n, v	m̩, n
n	ŋ, w	m	m m, l	l, m	ɾ̃, n̩, ŋ, m
ŋ	n, j, ɲ	(not tested)	(not tested)	(not tested)	n, m
ɲ	ŋ, w, j	(not tested)	(not tested)	(not tested)	(not tested)
r	w	(not tested)	n, b, v b, w	(not tested)	ɔˠ, ɜˠ
f	(none)	θ, k, s, p	p, h, b, θ, ð p, b, k, h, θ, v, ð	v	s, θ, v, z
v	w	ð, b, z	b, ð, h, f, θ b, f, ð, p, h, w, θ	b, g	f, ð, z, b
			p, b, f, ð, h, t		
ð	(not tested)	v, z, g, b	θ, l, b, v, n, z, d b, θ, l	(not tested)	θ, d, v, f, b
s	(none)	θ,ʃ, f	θ, ð, f, z θ, z, f, ð	f, ʃ	ʃ, z, f
z	dʒ	ʒ, g, ð, d, v	ð, θ, v, w ð, θ, b, n	(not tested)	s ʒ, v
ʃ	tʃ	(none)	tʃ tʃ	(none)	tʃ, ʒ, s
ʒ	(not tested)	z	(not tested)	(not tested)	ʃ, z, dʒ, tʃ, s, ʉ, n̩
h	k, j, p	(not tested)	p, f, t, k p, k, f, b, θ, v, t	(not tested)	ɦ, q, f
j	l	(not tested)	d dʒ, n	(not tested)	(not tested)
l	r	(not tested)	m, b, n, ð m, b, p, w	n	l̩, oʊ, w
w	j	(not tested)	m dʒ, b	(not tested)	l, ɔ, uː

Looking generally at each phoneme in the study, the /p/ was always confused with /f/ across different research stud-ies, only the frequencies of it occurring over other phonemes were different. Other than that, /p/ was also confused with /b/, except for Miller et al. [1]. /p/ was also constantly confused with /k/ except for the Malay study.

From the list of confusions happening across different language settings and experiments, there was almost no clear correspondence across languages. For the Malay confusion study, it was expected that the recognition rate was higher for consonants in the CV structure and especially less confusing for plosive and sounds originating between dental and post-alveolar due to the place combined with the manner of

articulation. This however was not proven. Based on general observation, it is believed that the voiceless phonemes tend to create more confusion than the voiced phonemes.

Cutler et al. [6]'s respondents tend to misidentify plosive phonemes as /h/. It can be easily dismissed as a technical error. However, the sound produced after post-alveolar onwards (towards glottal) may have been confused as /h/ due to the aspiration effect. This was indirectly supported by Miller et al. [1] quoted by Fant et al. [2], who stated that nearly all the confusions were in terms of different places of articulation with in a subclass of constant manner of articulation. From Miller et al. [1]'s 0 SNR feedback however, it was mostly true for the highest confusion in the list. For example, the phoneme /g/ was mainly confused as /d/ which was also a plosive but other confusions were from fricatives. For the phoneme /b/ however, the highest confusion was the phoneme /v/ which is a fricative but has a similar place of articulation. It sufficed to re-iterate that other studies used human voices with either added noise or re-sampled voiced.

According to Fant et al. [2], when the sounds were re-filtered, the feedback showed the clear tendency that dental stops and fricatives were almost never recognised as such. This is also similar in the Malay study. When confusions happened during the dental of plosive, fricatives or nasal, the frequency of the confusions will be more on the post-retroflex sounds. The phoneme/t/ was confused as the affricate /ʤ/, /n/ as /ŋ/ and /z/ as /ʤ/. Although /ʤ/ is not even a plosive, the sound is closely similar to the sound /ʃ/ which does not exist in Malay. Hearing/z/ as an affricate was understandably due to the burst of sound before /z/. These similarities were believed to happened because the synthesised speech was generated using a Malay speech synthesiser using the HTS approach. The recording was done at the 44 kHz but then during feature extraction, it was down sampled to 22 kHz. This could lead to some respondents hearing the sound as if it had been filtered.

The confusions were less prominent for CV syllables compared to VC syllables. The confusions also occurred less often when paired with vowel /a/ and /i/ rather than /ə/ and /e/. This also explained why the previous studies always used /a/ or /ɐ/. Confusion studies on VC syllables did not have a lot of comparable studies. Therefore the comparison study will be done on the coda consonants for VC and CVC syllable for Malay studies.

3.2 Phoneme Confusion Matrix for Consonants in Syllable VC

For the VC study, three vowels were paired to the Malay consonants. However, the pair of /ij/ was dropped because, it never occurs in Malay and following the standard Malay spelling, the /j/ sound is subtly assimilated into /i/. At the beginning of the experiment, it was believed that the VC syllable would produce more confusions than the CV syllable. From the respondents' feedback, it was more accurate to conclude that the confusions were sparser than the syllable structure CV as shown in Tab. 3. Some confusion also had higher frequencies than the phoneme itself. This indicated that some phonemes could be easily confused with others when the phonemes were at the coda position. It was also found that the voiced and voiceless phonemes had more confusion phonemes than consonants. However, the voiceless phonemes' confusions were caused by an individual's perception rather than the perceptual confusion itself. This can be observed by the frequencies of occurrences for some confusion phonemes.

For plosive phonemes, /p/ was not able to be identified as itself more than half of the occurrences. It was confused with /f/, /k/, /b/ and /v/. This was similar to what has been presented by Pinto et al. [4] where the /p/ was also confused as /k/, /f/ and /b/ in Pinto et al. [4] study. The similarity existed for Malay onset consonants where the /f/ and /b/ were also listed as confusion phonemes. The /b/ in Malay had a high identification as itself, but also was confused as /m/, /p/ and /v/. This again was similar to Pinto et al. [4] - /v/ and /p/ and Malay onsets - /v/ and /m/. For phoneme /t/, it was confused as /d/, /b/ and /k/ but the only similarity with Malay onsets was /d/ while Pinto et al. [4] also listed /d/ and /k/ as its confusions.

Comparing the response with the syllable CVC, the phoneme /t/ was confused as /d/ and /p/. This is similar to the feedback presented by Pinto et al. [4]. For the phoneme /d/, it was mainly confused with

/m/, /n/, and /t/. All belonged to dental. The phoneme /k/ was confused with /g/ however the phoneme /g/, other than being confused with /k/, was also confused with /ʤ/.

Table 3: Phonemes confusion for coda consonants for syllable structure: VC

Speech Synthesiser's Phones Production	Observed Phoneme Identified by Listeners																					
	b	tʃ	d	f	g	h	ʤ	k	l	m	n	ŋ	ɲ	p	r	s	ʃ	t	v	w	j	z
b	26			2						**10**				**7**				1	**6**	2		
tʃ		20			**4**	2	**23**										**4**	1				
d	3		17		2		1	3	2	**9**	**7**	2		1				**5**			2	
f	2			26		2	1		3		1				2		**5**	2	**8**	1		1
g			2		26	1	**4**	**7**	3	1	3	1		2	1			3				
h			1	2		19	**4**	**5**	2			1		1		**7**	3	**5**			**4**	
ʤ		**7**			3		43				1											
k		1			**9**	2	1	36	1		1			1				1			1	
l								1	28					**5**	**7**			2			**11**	
m	1									51	1			1								
n									1	**6**	40	**5**		1							1	
ŋ					2					**12**	6	32	1								1	
ɲ							1				**13**	**17**	21								2	
p	**7**			**12**	3			**8**	1	1		1		14				2	**5**			
r			1	2				2	**12**		1				27	1		**6**			2	
s											1					46	1					**6**
ʃ		**6**					1	1								3	43					
t	**4**	1	**6**	2			3	**4**	2					3	1			25			1	2
v	2			**11**	1				1	**11**		2			2				21	3		
w						2			3										3	44	2	
j	1					3			2											1	29	
z							**6**									**5**	3	1				39

The two affricates /tʃ/ and /ʤ/ were mostly confused with each other. The phoneme /tʃ/ was highly confused as /ʤ/. The confusion number was higher than the recognition of the phoneme itself. The phoneme /ʤ/ mostly was identified as itself. When confusions occurred, it was mainly perceived as /tʃ/. In real usage, these affricates rarely occur at the coda position in Malay words. When it did happen however, most of the time it was a /ʤ/. Examples of such usage are: *majlis* (ceremony), *majmuk* (plural), *buruj* (constellation), *hijrah* (migration), *koc* (train coach) and *Mac* (March).

Fricatives /f/ were confused as /v/ or /ʃ/. The confusions were quite scattered but two were the prominent ones. The /v/ confusions were also scattered but when confusions oc-curred, it was mainly detected as /f and m/. These were also true for the CVC syllable structure in the coda position. The /s/ was only confused as /z/ while /z/ was confused as /ʤ/ and /s/. The phoneme /S/ was confused as /tʃ/ instead of /s/ in the initial assumption. The confusions of /h/ were very scattered. It was mistaken as /s/, /t/, /k/, /ʤ/ and /j/.

The bilabial nasal had no confusions even at the coda position. However, /n/ was sometimes confused as /m/ or /ŋ/. /ŋ/ was mistaken as /m/ quite frequently and sometimes as /n/ while /ɲ/ had many confusions with /n/ and /ŋ/. This may be due to /ɲ/ almost never occurring in coda position in Malay.

For /r/, it was mainly confused as /l/ and as /t/. Both are dental. For other glides: /w/ and /j/, no prominent confusion occurred.

3.3 Phonemes Confusion Matrix for Onset in Syllable CVC

It was expected that the observation on the onset of a CVC syllable would show consistency in phoneme confusions given a better context (due to the adjacent consonants to the vowel). From Tab. 4, the distribution can be seen as less sparse than in Tab. 1. It is believed that this is due to the structure of the syllables, the respondents being surer of what they thought they heard and thus perceiving less ambiguity.

As with the CV structure, the phoneme /p/ was also mistaken as the phonemes /f/ and /b/. But the confusions were heavily focused on /f/ and the same with /b/ where confusions were heavily focused on /v/. However, a few phoneme confusions identified with /d/. For the phoneme /t/, the confusions were

only with /ʧ/ and /d/ which showed noticeable reduction of confusions as compared to the syllable CV. The phoneme /d/ was not confused much other than /t/. The phoneme /k/ was confused as /ʧ/. The similarity between the two were that both have 'plosiveness' as manner of articulation. The phoneme /g/ also had multiple confusions like the CV structure. It was confused as /ʤ/, /d/ and /k/.

Table 4: Phonemes confusion for onset consonants for syllable structure: CVC

Speech Synthesiser's Phones Production	Observed Phoneme Identified by Listeners																				
	b	ʧ	d	f	g	h	ʤ	k	l	m	n	ŋ	ɲ	p	r	s	ʃ	t	v	w	j	z
b	29		**4**																**19**	2		
ʧ		44					**9**										1					
d	3		36		1		2		3		1	1			1			**4**	1			1
f				46												3			**5**			
g			**5**		33		**8**	**5**										2				1
h		1		3		44		1						**4**							1	
ʤ							54															
k		**7**		1	1	2		43														
l									50		1										3	
m				1						45	1	1								**6**		
n			1						**6**	2	35	**6**									**4**	
ŋ									3	2	**5**	31	2							1	**10**	
ɲ											3	**9**	37								**5**	
p	**4**			**18**										28		1		3				
r				3											45				**5**	1		
s		1														53						
ʃ		**6**					1									2	45					
t		**6**	**5**				2									1		40				
v									1									1	51	1		
w															2				2	48	2	
j									**9**						1	1					43	
z							1															53

The affricate /ʧ/ was confused as /ʤ/; however /ʤ/ was not confused at all at the onset of syllable CVC.

For fricatives, when confusion happened for the phoneme /f/, it was perceived as /v/. The phoneme /v/ however, was not mistaken at all. A similar condition was found for the phonemes /s/ and /z/. The phoneme /ʃ/ was confused as /ʧ/, while the phoneme /h/ was mistaken as /p/.

The nasal confusions were less consistent for consonants at the onset of a CVC structure as compared to CV. As with the confusions in CV for /m/, it was also mistaken as /w/ but with lesser frequency. The phoneme /n/ was mistaken as /l/, /ŋ/ and /j/. The phoneme /ŋ/ was very sparsely distributed but has the consistency of being mistaken as the phonemes /n/ and /j/. Finally, /ɲ/ was confused as /ŋ/ and /j/.

The phoneme /r/ was sometimes mistaken as /v/. The phoneme /j/ was mistaken as /l/, however the phonemes /l/ and /w/ were not mistakenly perceived at all.

It can be concluded that at the onset position, when more phoneme s were provided for the respondent to guess, the phoneme was less likely to be mistaken as another phoneme . However, it can also be observed that some phonemes can really be confused as something else and multiple phoneme s were confused as its voiceless/ voiced pair. It also happened due to the vowel used in the pair as well as the second consonant (its coda) usage.

3.4 Phonemes Confusion Matrix for Coda in Syllable CVC

When constructing the CVC syllables for the confusion study, the focus was specifically on the coda of the syllable and the vowel usage. Before obtaining the results for phoneme confusions at the coda of the syllable CVC, it was assumed that the confusions would closely reflect the phoneme confusions at the coda for syllable VC as shown in Tab. 3. However it was later found that this was not exactly true as what is concluded in Tab. 5.

Table 5: Phoneme confusion for coda consonants for syllable structure: CVC

Speech Synthesiser's Phones Production	Observed Phoneme Identified by Listeners																					
	b	tʃ	d	f	g	h	dʒ	k	l	m	n	ŋ	ɲ	p	r	s	ʃ	t	v	w	j	z
b	6			**4**	1	2		1		**11**	**7**	2		**10**				**7**	2	1		
tʃ		28			1		**18**										**5**	2				
d	2		19	1		1	1	**4**		**5**	**4**	1		**4**	2	1		**8**	1			
f				35		**6**		3						1		3		2	**4**			
g	3		1	3	28			**10**			1	1		3				2	1		1	
h				**6**	1	34		**5**	1						1			3			3	
dʒ		**13**			2		37				1							1				
k	1				3			40			1			2				**6**			1	
l				1					13		**19**	**4**			**6**				1	2	**8**	
m										49	**4**										1	
n									2	2	35	**15**										
ŋ				1						**17**	**11**	24		1								
ɲ											**29**	**8**	14		1						2	
p	**4**		1	**10**		2		**5**		1	1			25				**5**				
r						3	1		**13**		1	2			27	2		1				**4**
s															1	48						**5**
ʃ		3														**4**	45					2
t	3		**13**	3	2			3	1					**8**				21				
v	3			**13**	2	2				**8**	1				1	**6**			**17**	1		
w									1			1	1							41	**10**	
j				1		2			**6**									1			44	
z			1				2									**8**	4					39

The summary of confusions between VC and CVC is presented in Tab. 6 in the respective columns. There was no indication that CVC or VC adds context to the articulation sequence that helps with the identifications. However, it can be seen that the confusions between CVC and VC for phoneme /p/, /b/, /t/ and /d/ are similar across different languages and also for /n/, /ŋ/, /ɲ/ and /r/. For the first case it is due to the frontal nature of sound and for the later is because of the nasalised and trill effect. It can also be observed that the glides and liquids tend to be confused with each other. It should be emphasised that /r/ is not consistently a trill in Malay. When a very similar pronunciation is produced, Malay native speakers will easily accept it as an "r", despite it being produced as /ɾ/, /ɽ/ or sometimes to the extent of /ʁ/ (which might happened due to lack of practise of trill or tap during childhood). For synthesised speech, the sound may be produced as /r/ or /ɾ/ because it was what was being produced by the training voices.

Because /k/ and /g/ plosiveness originates from uvular and differs only in the voiced/voiceless categories, it was assumed they will be confused with each other. However, it was not true for /k/ which was where confusions happened; it was mainly perceived as /t/. /tʃ/ and /dʒ/ were again confused with each other although /tʃ/ confusions with /dʒ/ were higher than vice versa. Fricatives tended to be confused with its voiced or voiceless pair.

Despite being different, there were slight patterns of confusions that can be seen across languages. For fricative confusions, since the affricates existed in the languages (as listed in Tab. 6), respondents tended to confuse the fricatives as affricates or the corresponding plosive counterpart of the affricates besides the phoneme 's own neighbour.

In the represented study, Malay has six phonemes: /a/, /e/, /ə/, /i/, /o/ and /ʊ/. However, there were more sounds due to style of talking, dialects and influence from first language. For example, if the "a" sounded like /a/, /ɐ/, /æ/ or /ʌ/, it would still be understandable and written as "a". There were also confusions for the phoneme /e/. The /e/ also usually produced as /ɛ/ or sometimes /ɜ/. This study skiped the vowel's confusion phoneme for Malay. This was due to the looseness of vowel pronunciation, and non-systematic writing system in Malay which made it problematic to distinguish such occurrences except for those respondents familiar with the IPA writing system. Based on the study by Bhatara et al. [19], second language skills may sometimes interfere with emotion recognition from speech prosody, particularly for positive emotions. Similar to Paone et al. [20], also suggest that learners' L2 knowledge might contribute

to the ability to infer emotions from L2 speech prosody and to judge the intensity as well as native speakers do. This is also the reason why logatome utterance are being used in this evaluation although most respondents are at least sufficiently proficient in their second or third language.

Table 6: Phoneme confusion comparison for coda consonants for syllable structure: VC and CVC

Phoneme	Malay VC	Malay CVC	Cutler et al. (2004) - English	Cutler et al. (2004) - Dutch
p	f, k, b, v	f, k, t, b	t, k, f, θ	b, t, k, θ, f, d
b	m, p, v	m, p, t, n, f	v, d, ð, k	d, v, t, θ, ð, p
t	d, b, k	d, p	k, p, θ	d, θ, k, ð, p
d	m, n, t	t, m, k, n, p	v, ʤ, n, ʒ, g	t, ð, θ, ʤ
k	g	t	t, p	t, g, p, θ, f
g	k, ʤ	k	d, b, ð, θ	d, t, ʤ, ð, k
ʧ	ʤ, ʃ, g	ʤ, ʃ	ʤ	ʤ, ʃ, ʒ
ʤ	ʧ	ʧ	ʒ, d	ʧ, ʒ, d, ð
m	(none)	n	ŋ, n, v	n, t, ŋ, dθ
n	m, ŋ	ŋ	m, ŋ, d	t, ŋ, m, d
ŋ	m, n	m, n	n, m, g, v	n, m, t, g, d
ɲ	ŋ, n	n, ŋ	(not tested)	(not tested)
r	l, t	l, z	(none)	t, d
f	v, ʃ	h, v	θ, p, t, k, ð	t, d, θ, p, k, ð
v	f, m	f, m, s	f, g, ð, d	d, t, f, ð, θ, b, l
θ	(not tested)	(not tested)	f, t, ð, p	t, f, ð, d, p
ð	(not tested)	(not tested)	d, v, ʤ, z, ʒ, g	d, t, v, θ, ʤ
s	z	z	f, θ	f, θ, ð, ʃ, z
z	ʤ, s	s, ʃ	v, d, ð, s, ʒ, ʤ	s, θ, ð, d, v, ʒ
ʃ	ʧ	s	ʧ	ʒ, s
ʒ	(not tested)	(not tested)	ʤ, ð, v	ʃ, ʤ, z, ʧ, ð
h	s, k, t, ʤ, j	f, k	(not tested)	(not tested)
j	(none)	l	(not tested)	(not tested)
l	j, r, p	n, j, r, ŋ	f, v	d, t, f, r
w	(none)	j	(not tested)	(not tested)

Since consistent confusions were difficult to obtain, a more direct approach to the confusions survey was conducted. Given a specific context, respondents were asked to listen and type back what they heard from the list of sounds.

4 Intelligibility on Substituted Phoneme's Words

Mix of valid words with substituted phoneme s were evaluated together with words which had not undergone any changes. The words with change of phonemes were added into the pronunciation dictionary to ensure that the intended sounds were produced. Then, a call for respondents was made to evaluate the intelligibility of the sound. The surveys were run based on the assumption that the phoneme s could be substituted with another phoneme in certain conditions so as to imitate the original word pronunciation. The intelligibility tests were conducted by letting the respondents run the survey at their own convenience and pace. All respondents conducted the survey using a pair of headphones.

Seventeen respondents participated in the survey with no specific language background and age between 18 to 50 years old. However, all can speak Malay as the first or the second language. From the evaluation, there were 124 conditions where the modified words with one modified phoneme were perceived as the intended words. It is important to state that the modified words becoming invalid words when the changes made, and the respondents were already advised to write what they think they heard. This experiment was conducted to identify how the respondents perceived the synthesised speech in general. There were 111 correctly identified words and 50 incorrectly identified words (from the controlled sample).

The sensitivity of the overall feedback was 0.7164. The misclassification was 0.2835. To further analyse the results, a test of statistical significance was conducted using the chi square test.

The value of chi square, $\chi 2$ was 1.1366. However for degrees of freedom (df) equal to 1 and $P = 0.05$, $\chi 2$ must be equal or exceed 3.84 to be significant. Therefore in terms of intelligibility testing, perceiving the substituted phoneme as the intended word was possibly due to chance.

5 Perception Based on Context

To further evaluate the possibility of that such a substitution can be perceived as the intended sound, a set of perceptual tests was conducted on onset and coda modifications given a context evaluation. In this evaluation, a string of three or four phoneme s was arranged in sequence and each utterance had similar rhyme. One of the utterance would have a slight change: either the onset or coda was different from the others. These experiments were conducted on the assumption that it was easier for the respondents to confuse the sound of the different onset or coda due to the neighbouring words. The respondents were simply told to type back what they heard. There were 24 respondents in the study and all can speak in Malay and aged between 25 to 55 years old.

In the onset evaluation, 138 responses identified words that that were affected by the neighbouring words and 81 words were not. For the control set of sounds, in the total of 519 words paired into three to four words sequences, 395 words were correctly identified by respondents and 124 were not. Among the data (utterance) that underwent phoneme substitution, the sensitivity was 0.6301. The specificity of the study when there was no modification of the phoneme of the words was 0.7611. The overall sensitivity was 0.7222 and the misidentification was 0.2778.

A test of statistical significance was conducted. The value of $\chi 2$ was 13.1627. For $df = 1$ and $P = 0.05$, $\chi 2$ must equal or exceed 3.84 to be significant. Therefore it can be said that for phoneme s substituted with matching phoneme s in a specific context, the perception will be affected by the neighbouring words and is statistically significant for onset consonants replacement.

As with onset evaluation, the assumption was that when the coda of a word was substituted with a similar phoneme in a selected context, the perception will be affected by the neighbouring words. It was hypothesised that if such a condition happened, it should not happen due to chance. For synthesised speech, respondents identified 98 words that were affected by the neighbouring words and 76 words that were not. The control words (no modification made to those words) found that 353 words were correctly identified, and 113 were not.

Among the data (words) that underwent phoneme substitution, the sensitivity was 0.5632. The specificity of the study for control words was 0.7575. In total, the overall sensitivity was 0.7047 and misclassification was 0.2953.

The value of $\chi 2$ was 22.9820. For $df = 1$ and $P = .05$, $\chi 2$ must equal or exceed 3.84 to be significant. Therefore it can be said that for coda phoneme s substituted with matching phoneme s in a specific context, the perception will be affected by the neighbouring words and is statistically significant.

These experiments were also conducted using synthesised speech. The feedback was expected to be also influenced by the machine generated speech and therefore the expected sensitivity and specificity were better than expected.

This showed that the confusion phones presented in Section 3 are applicable for use in phoneme substitution as long as they occur with in the phoneme range listed in the confusion list. From the value of $\chi 2$ for both onset and coda, it was believed that the coda might be better accepted as a substitution than as an onset where the changes (in the coda) were less frequently detected by the respondents.

6 Conclusion

The issue investigated were of reusing other resources to create another TTS albeit with the substantial chance of not having complete data, in particular the trained phoneme s used in the resource language. The possibility of obtaining a substitute was investigated because it eliminated the need of new training or recording, as suggested by Kominek [21]. Without such, it is certain that synthesised speech will not sound native in the best possible situation and is distorted to be unintelligible in the most undesirable scenario.

To avoid the worst-case scenario, a study on possible substitutes was conducted via the study of confusion matrix. By completing the confusion matrix, it is hoped that the best replacement candidate can be identified. Based on the findings from this study, intelligibility and perception tests based on simple listening and a contextual perception listening test were conducted. From the intelligibility test, the phoneme susbtitution is not suitable to be apply when the pronunciation is to be done at an individual lexicon level, for example the pronunciation provided by online dictionaries. For such pronunciation, if phoneme substitution is used, the sound produced will be noticeably distorted. The results also showed that for perception evaluation, the onset modifications can be perceived as intended as long as the context were given. The results obtained were tested for significance testing, where the perception listening test being found to be statistically significant. It shows that phoneme substitution is possible if conducted with carefully selected substitutions given in a meaningful context.

Funding Statement: This manuscript publication fee is partially funded by the Universiti Sains Malaysia's Short term grant no: 304/PKOMP/6315273.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] G. A. Miller and P. E. Nicely, "An analysis of perceptual confusions among some English consonants," *Journal of Acoustic Society of America*, vol. 27, no. 2, pp. 338–352, 1955.

[2] G. Fant, B. Lindblom and A. de Serpa-Leitao, "Consonant confusions in English and Swedish. A pilot study," *Speech Transmission Laboratory. Quarterly Progress and Status Reports Journal*, vol. 7, no. 4, pp. 31–34, 1966.

[3] A. Lovitt and J. B. Allen, "50 years late: Repeating miller-nicely 1955," in *Proceeding of INTERSPEECH*, 2006, pp. 2154–2157, 2006.

[4] J. P. Pinto, A. Lovitt and H. Hermansky, "Exploiting phoneme similarities in hybrid HMM–ANN keyword spotting," in *Proc. of the 8th Annual Conf. of the International Speech Communication Association (INTERSPEECH 2007)*, Antwerp, Belgium, pp. 1610–1613, 2007.

[5] B. T. Meyer, M. Wächter, T. Brand and B. Kollmeier, "Phoneme confusions in human and automatic speech recognition," in *Proc. of the 8th Annual Conf. of the International Speech Communication Association (INTERSPEECH 2007)*, Antwerp, Belgium, pp. 1485–1488, 2007.

[6] A. Cutler, A. Weber, R. Smits and N. Cooper, "Patterns of English phoneme confusions by native and non-native listeners," *Journal of Acoustic Society of America*, vol. 116, no. 6, pp. 3668–3678, 2004.

[7] T. U. Christiansen and S. Greenberg, "Perceptual confusions among consonants, revisited—Cross-spectral integration of phonetic-feature information and consonant recognition," *IEEE Transaction on Audio. Speech and Language Processing*, vol. 20, no. 1, pp. 147–161, 2011.

[8] P. Karanasou, L. Burget, D. Vergyri, M. Akbacak, A. Mandal *et al.*, "Discriminatively trained phoneme confusion model for keyword spotting," in *Proc. of 13th Annual Conf. of the International Speech Communication Association (INTERSPEECH 2012)*, Portland, Oregon, USA, pp. 2434–2437, 2012.

[9] A. Žgank, B. Horvat and Z. Kačič, "Data-driven generation of phonetic broad classes, based on phoneme confusion matrix similarity," *Speech Communication Journal*, vol. 47, no. 3, pp. 379–393, 2005.

[10] A. Leijon, G. E. Henter and M. Dahlquist, "Bayesian analysis of phoneme confusion matrices," *IEEE/ACM Transaction on Audio, Speech, and Language Processing*, vol. 24, no. 3, pp. 469–482, 2015.

[11] L. -F. Shi and N. Morozova, "Understanding native Russian listeners' errors on an English word recognition test: Model-based analysis of phoneme confusion," *International Journal of Audiology*, vol. 51, no. 8, pp. 597–605, 2012.

[12] B. Ranaivo and N. -H. Samsudin, "Bahasa Malaysia phoneme s," Internal Report, Computer Aided Translation Unit, Universiti Sains Malaysia, 2003.

[13] MBROLA-Group, "The MBROLA projects: Towards a freely available multilingual speech synthesizer," Mons, Belgium, 2005, [Online]. Available: http://tcts.fpms.ac.be/synthesis/mbrola.html.

[14] H. Li, M. Aljunied and B. S. Teoh, "A grapheme to phoneme converter for standard Malay," in *Proc. of Oriental Commitee for the Coordination and Standardisation of Speech Databases and Assessment Techniques (O-OCOSDA)*, Jakarta, Indonesia, 2005.

[15] Z. Ahmad, N. H. Jalaluddin, F. M. M. Sultan, H. Radzi, M. S. Yusof *et al.*, "Pemerkasaan jati diri bahasa melayu: Isu penyerapan kata asing translated as empowerment and configuration malay language identity: Infiltration of foreign vocabularies issue," *Jurnal Melayu*, vol. 6, pp. 13–27, 2011.

[16] IPG, "Sejarah perkembangan bahasa melayu, perkamusan dan terjemahan," Kementerian Pendidikan Malaysia, Dewan Bahasa Pustaka, pp. 133–158, 2011.

[17] A. Clyness and D. Deterding, "Standard malay (Brunei)," *Journal of International Phonetic Association*, vol. 41, no. 2, pp. 259–268, 2011.

[18] S. Bhatt, A. Dev and A. Jain, "Confusion analysis in phoneme based speech recognition in Hindi," *Journal of Ambient Intelligence and Humanized Computing*, vol. 11, no. 10, pp. 4213–4238, 2020.

[19] A. Bhatara, P. Laukka, N. Boll-Avetisyan, L. Granjon, H. A. Elfenbein *et al.*, "Second language ability and emotional prosody perception," *PLOS ONE Journal*, vol. 11, no. 6, pp. 1–13, 2016.

[20] E. Paone and M. Frontera, "Emotional prosody perception in Italian as a second language," in *10th Int. Conf. of Experimental Linguistics*, 2019.

[21] J. Kominek, "TTS from zero: Building synthetic voices for new languages," *Ph.D. dissertation*, Language Technologies Institute, School of Computer Science, Carnegie Melon University, Massachusetts, 2009.