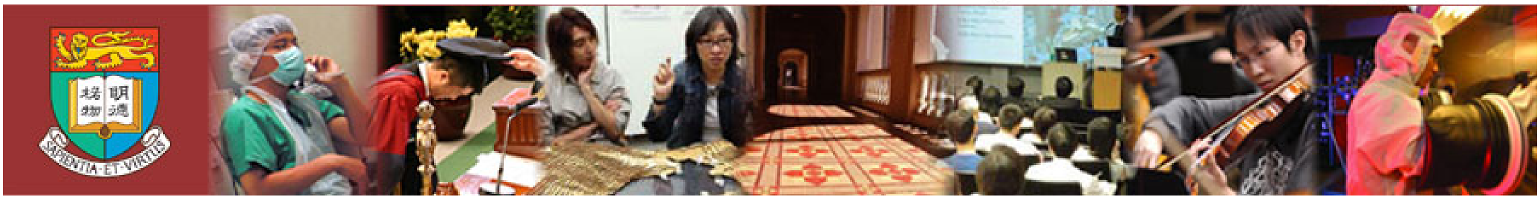

Perceptual contribution of vowels to Mandarin sentence intelligibility under conditions of spectral degradation

Wong Wai Kwan

A dissertation submitted in partial fulfilment of the requirements for the Bachelor of Science (Speech and Hearing Sciences), The University of Hong Kong, June 30, 2014.

Perceptual contribution of vowels to Mandarin sentence intelligibility under conditions of spectral degradation

Wong Wai Kwan

Abstract

A recent study showed a vowel advantage over consonant to sentence intelligibility in Mandarin. Considering the fact that many important acoustic cues for sentence intelligibility are contained in the vowel segment, the present study investigated the effect of spectral degradation and its interaction effect with vowel duration on Mandarin vowel-only sentence intelligibility. Three types of spectrally degraded stimuli, including fundamental frequency flattened (*F_0F*), sine-wave synthesized (*SWS*) and noise-vocoded (*NV*) vowel-only sentences, were generated. Different proportions of vowel centers were preserved by using a noise-replacement paradigm. Listening experiments showed that fundamental frequency contour only had a minimal effect to vowel-only sentence intelligibility, while harmonic cues had a more notable effect. Intelligibility of *NV* sentences was significantly lower than that of *SWS* sentences, suggesting other acoustic cues such as formant frequency information contribute to the vowel advantage when harmonic cues are discarded. Discarding vowel edges had a significantly negative effect on vowel-only sentence intelligibility under conditions of spectral degradation. The present study supports emphasis on the preservation of harmonic cues and vowel duration in speech processing strategies.

1. Introduction

Many studies based on noise-replacement paradigm, which replaces vowel (*V*) or consonant (*C*) segment by noise, found that *V*-only sentences yielded a 2:1 sentence intelligibility advantage over *C*-only sentences in English (e.g. Kewley-Port, Burkle, & Lee, 2007). Similarly, a *V* advantage over *C* to sentence intelligibility was found in Mandarin, and even yielded a more significant 3:1 intelligibility ratio (Chen, Wong, & Wong, 2013). Frequency and amplitude are two main acoustic features that can be found in any natural speech signals. Fundamental frequency (F_0) contour, formant frequencies, temporal envelope (amplitude variation across time) and harmonics are examples of fine acoustic cues. With the *V* advantage to sentence intelligibility in English, many studies investigated the contributions of acoustic cues to account for the *V* advantage. For example, higher sentence intelligibility was found with greater resolution of temporal envelope in the vowel portions when consonants were preserved (Fogerty, 2012); envelope cues were found to be responsible for the context-dependent advantage of *V*, whereas dynamic F_0 cues were not (Fogerty & Hume, 2012).

However, little has been done to examine the acoustics cues that contribute to the *V* advantage in Mandarin sentence intelligibility. A need to investigate this issue specific to Mandarin is supported by the greater *V* advantage found in Mandarin than in English, which may imply language-specific effects (Chen et al., 2013). Mandarin and English are different in many aspects. One of the most apparent differences is that Mandarin is a tonal language, in which tone information is essential to lexical contrasts, whereas English is a non-tonal language. The four lexical tones in Mandarin, including high, rising, mid-falling and then rising, and high-falling tones, convey different meanings for the same phonemic structure. Dissimilarly, no tonal differentiation but different combinations, structures and number of syllables convey different lexical meanings in English. Tone was found to be as important as *V* and *C* to sentence recognition in the presence of temporal envelope cues in Mandarin (Fu, Zeng, Shannon, & Soli, 1998). The acoustic cues important for tone identification, including F_0 contour, amplitude variations and vowel duration, are mainly contained in the vowel segment (Chen & Loizou, 2011; Howie, 1976). Therefore, these acoustic cues, of which some are seemingly unimportant in accounting for the *V* advantage in English, may account for the *V* advantage in Mandarin due to their contribution in tone perception. On the other hand, Chen, Wong, and Hu (2014) recently showed that tonal information was relatively redundant for Mandarin sentence comprehension under quiet conditions, as high sentence recognition scores (94% and 95% respectively) were resulted from flattened- and randomized-tone sentences. Similar findings were also derived from Feng, Xu, Zhou, Yang, and Yin (2012), which found a mean recognition score of 91.6% in Mandarin sine-wave sentences despite a chance-level tone recognition performance. The contrasting results from the above studies put the roles of acoustic cues conveying tonal information for Mandarin sentence intelligibility in doubt.

Moreover, top-down processing is an important factor in considering sentence intelligibility. Listeners may use their semantic and syntactic knowledge of the language to derive meaning of the lost information from other lexicons in case of breakdown, which often happens in daily communication such as listening in noisy environment. Spectral cues such as F_0, envelope cues and harmonics might have contributions to the context-dependent processing of sentences, as in English.

In addition to different acoustic cues contributing to sentence recognition, temporal cues, such as vowel duration, syllable duration and duration of *V-C* boundary transition, contribute to tone and sentence recognition in Mandarin (Chen et al., 2013; Whalen & Xu, 1992). For instance, in Chen et al. (2013), Mandarin sentence intelligibility decreased significantly from 97.4% to 89.6% when the preservation of vowel centers reduced from 80% to 60%. Therefore, it is worthwhile to investigate if there is any interaction between the acoustic cues contained in vowel segment and vowel duration to sentence intelligibility.

It is worth to note that acoustic cues in vowel segment have been found to contribute more to sentence intelligibility even when *V-C* transition and difference in durations of *V* and *C* were taken into consideration (Fogerty & Kewley-Port, 2009). Together with the high *V*-to-*C* advantage in Mandarin sentences mentioned, we are motivated to investigate the above effects in *V*-only Mandarin sentences.

1.1. Purpose of the present study

The aim of the present study is to investigate the effects of spectral degradation and vowel duration on the intelligibility of Mandarin *V*-only sentences. Three types of spectrally degraded stimuli, processed by flattening F_0 (F_0F) contour, sine-wave synthesis (*SWS*) and noise-vocoding (*NV*) respectively, were generated. It is expected to establish conclusions about the contributions of different acoustic cues (i.e. F_0 contour, harmonics and formant structures) to *V*-only sentence intelligibility specific to Mandarin by comparing their effects on perceptual recognition of speech. As different degrees of spectral degradations are often experienced by hearing impaired patients, the present study might also give insights to the design of novel speech processing strategies for assistive hearing devices, such as hearing aids and cochlear implants.

In viewing the language difference between Mandarin and English, lexical tone information is unique to Mandarin. F_0 contour is the dominant cue for tone perception (e.g. Howie, 1976; Whalen & Xu, 1992). However, recent study has shown that this information does not lower sentence intelligibility much under quiet conditions, as tone information is compromised in natural speech (Chen et al., 2014). On the other hand, F_0 contour facilitates the prediction of syntactic units in sentences, which is useful in the top-down processing of sentences (Laures & Weismer, 1999). Thus, we hypothesize that the intelligibility would be reduced with flattened fundamental frequency vowel-only sentences (F_0F-V) compared to unprocessed *V*-only sentences with intact F_0 contour, but the effect would be little.

Sine-wave speech contains little tone information as no F_0 information is preserved (Feng et al., 2012). It also contains no harmonic information. This kind of processing greatly reduces acoustic information available compared to F_0F processing especially in terms of harmonic cues, which are important to speech recognition (Remez, Rubin, Berns, Pardo, & Lang, 1994). We predict that the intelligibility of sine-wave synthesized vowel-only sentences (*SWS-V*) would be notably reduced compared to that of F_0F-*V*.

In noise-vocoded speech, only temporal envelope cues, but no other temporal fine structures and harmonics are preserved in each noise band. Nevertheless, many studies (e.g. Chen & Loizou, 2011) showed that high intelligibility scores could still be achieved with an adequate number of channels used. Envelope cues could facilitate sentence comprehension as it was found to facilitate word prediction in English, and could therefore facilitate sentence perception (Waibel, 1987). As for *SWS* processing, we predict that noise-vocoded vowel-only sentences (*NV-V*) would have lower intelligibility than F_0F-*V* as it lacks harmonic information essential to speech perception, but a higher-than-chance intelligibility would still be yielded. Assuming significant effects of F_0 contour and harmonic cues on Mandarin *V*-only sentence intelligibility, we would also investigate if there is any difference between their relative extents of effect to *V*-only sentence intelligibility.

Both *SWS* and *NV* processing strategies limit the harmonic structure and F_0 information present. However, acoustic differences exist between their stimuli generated. For example, a little amount of formant information such as average frequencies and amplitudes is preserved in *SWS* speech, as it is composed of sinusoidal replicas of the first three formants; but all formant information is eliminated in *NV* speech. In this study, we would observe if any significant difference exists between intelligibilities of *SWS-V* and *NV-V* sentences, which might imply presence of other acoustic cues that contribute to the *V*-advantage over *C* in sentence intelligibility.

Increasing vowel duration has been shown to increase Mandarin sentence intelligibility (Chen et al., 2013). Similarly, we hypothesize that sentence intelligibility would increase with longer vowel duration in this study. We also predict that there would be an interaction effect between spectral degradation and vowel duration, given that the duration of acoustic cues for sentence recognition is decreased when vowel duration is reduced.

2. Method

Twenty young normal-hearing native Mandarin listeners, including seven male and thirteen female, aged 18 to 27 years old (*M* = 20.75) were recruited by convenient sampling from The University of Hong Kong. They were paid for their participation. Hearing screening was conducted to ensure their pure-tone air-conduction threshold of each ear was at or below 20 dB HL at octave frequencies from 250 to 8000 Hz (American National Standards Institute, 1996). None of the subjects had participated in studies listening to spectrally degraded sentences before. The above procedure is to ensure that the result of the present research is not affected by hearing loss or lack of language proficiency.

The sentence materials were adopted from the Mandarin Chinese version of the Hearing in Noise Test (*MHINT*) (Wong, Soli, Liu, Han, & Huang, 2007). Twenty-four lists from the database, each composed of ten ten-syllable Mandarin sentences were used. All the sentences were audiotaped by a male speaker with F_0 ranging from 75 to 180 Hz. Stimuli containing different amounts of spectral cues and vowel duration in the vowel segment were then generated. Three types of speech processing strategies were used for discarding spectral cues, as follows:

2.1 Speech processing 1 – flattening fundamental frequency (F_0F)

The dynamic F_0 cues of the sentence materials were extracted and then replaced by the mean value of F_0 of the utterance. The natural F_0 contour was therefore flattened, while other acoustic cues such as formant variations and harmonics were preserved.

2.2 Speech processing 2 – sine-wave synthesis (*SWS*)

In *SWS* processing, the center frequencies of the first three formants (F_1, F_2 and F_3) of the sentence materials were extracted and replaced by sinusoid replicas. Spectrograms of the generated *SWS* stimuli and those of the original sentences were compared to ensure the sine waves match the formant frequencies of the sentence materials. Through this process, only information about the center frequencies of F_1, F_2 and F_3 was preserved, while other acoustic cues (including F_0 and harmonics) were removed. Implementation of the above *SWS* (and F_0F) processing was performed by using a computer software PRAAT (http://www.linguistics.ucla.edu/faciliti/facilities/acoustic/praat.html#noisespeech).

2.3 Speech processing 3 – noise vocoding (*NV*)

In *NV* processing, the sentence materials were passed through a pre-emphasis high-pass filter with 2000 Hz cutoff and 3 dB/octave roll-off. They were then divided into six frequency bands between 80 and 6000 Hz by band-pass filters, as five to eight bands were shown to yield comparably high levels of intelligibility (Allen, 1994; Hill, McRae, & McClellan, 1968; Loizou, Dorman, & Tu, 1999). Amplitude envelope was extracted from each band by a full-wave rectification and a low-pass filter with cut-off frequency 160 Hz. The envelope signal extracted was used to modulate white noise, which was passed through the same band-pass filters again. The envelope-modulated noises of each band were summed up, and the level of the synthesized speech was adjusted to the root-mean-square value of the original speech. Through this process, spectral information was removed. Only temporal envelope cues and their co-varying information (i.e. brief information of F_0 contained in temporal envelope) were preserved.

Using the above three types of spectrally degraded stimuli and unprocessed stimuli, this study further generated *V*-only sentences by replacing segments other than the vowel portions with speech-shaped noise with signal-to-noise ratio -16 dB. *V* and *C* boundary information available in *MHINT* sentences were labeled manually by an experienced phonetician by observing acoustic landmarks on spectrograms using PRAAT, followed by verification by another experienced phonetician. The *V*-only sentence materials contained different amounts of vowel duration: preserving 100%, 80%, 60%, 40% or 20% of vowel centers. The rest of information including *C* and the remaining vowel portions from vowel onset and offset were eliminated by noise-replacement paradigm. A proportion factor f (f= 0.0, 0.1, 0.2, 0.3, and 0.4 respectively) was used to denote the proportion of vowel onset and offset durations to be replaced by noise. For example, f = 0.0 meant no vowel segment from onset and offset was to be replaced by noise; f = 0.1 meant 10% from vowel onset and 10% from vowel offset were replaced by noise, thus preserving 80% of vowel center.

A total of 20 testing conditions, composed of five different vowel center durations (f = 0.0, 0.1, 0.2, 0.3, and 0.4), and three types of spectral degradations (F_0F, *SWS*, and *NV*) and a control condition (i.e. unprocessed *V*-only sentence without spectral degradation), were used in the study. The experiment was carried out in a sound-proof booth at The University of Hong Kong individually for each participant. Stimuli were played through a circumaural headphone binaurally at a comfortable listening level to the listeners. Practice (i.e. with feedback) of 40 non-experimental sentences was given prior to the experimental conditions. Participants then participated in all the 20 experimental testing conditions, with each containing ten sentences. The order of the testing conditions was randomized and counterbalanced across listeners. No sentence was repeated within and across conditions. Participants were allowed to listen to each stimulus for three times in maximum, and required to repeat as many words as they could recognize. Their responses were scored and digitally recorded. A five-minute break was given for every 30 minutes.

3. Results

Sentence intelligibility score was calculated by dividing the total number of correctly recognized syllables by the total number of syllables in each testing condition. The mean sentence intelligibility scores are shown in *Figure 1* as a function of spectral degradation and vowel duration. The scores were converted to rational arcsine units (*RAU*) using the rationalized arcsine transform, which is linear and addictive, to stabilize the error variance (Studebaker, 1985). Note that all the following statistical analyses were conducted with scores in *RAU*. By using intelligibility scores in *RAU* as dependent variable, and spectral degradation and vowel duration as two within-subject factors, two-way repeated measure analysis of variance (ANOVA) was carried out. Mauchly's test indicated that the assumption of sphericity had been violated for the main effect of vowel duration [$\chi^2(9) = 19.64$, $p = .02$], and interaction effect of spectral degradation and vowel duration [$\chi^2(77) = 105.43$, $p = .03$]. Therefore, degrees of freedom were corrected using Greenhouse-Geisser estimates of sphericity ($\varepsilon = .69$ for the main effect of vowel duration; $\varepsilon = .51$ for the interaction effect of spectral degradation and vowel duration). The results showed significant main effects of spectral degradation [F(3,57) = 778.11, p < .001, η^2 = .98] and vowel duration [F(4,76) = 573.39, p < .001, η^2 = .97], and a significant interaction effect of the above two factors [F(12,228) = 25.13, p < .001, η^2 = .57] at significance level α = .05.

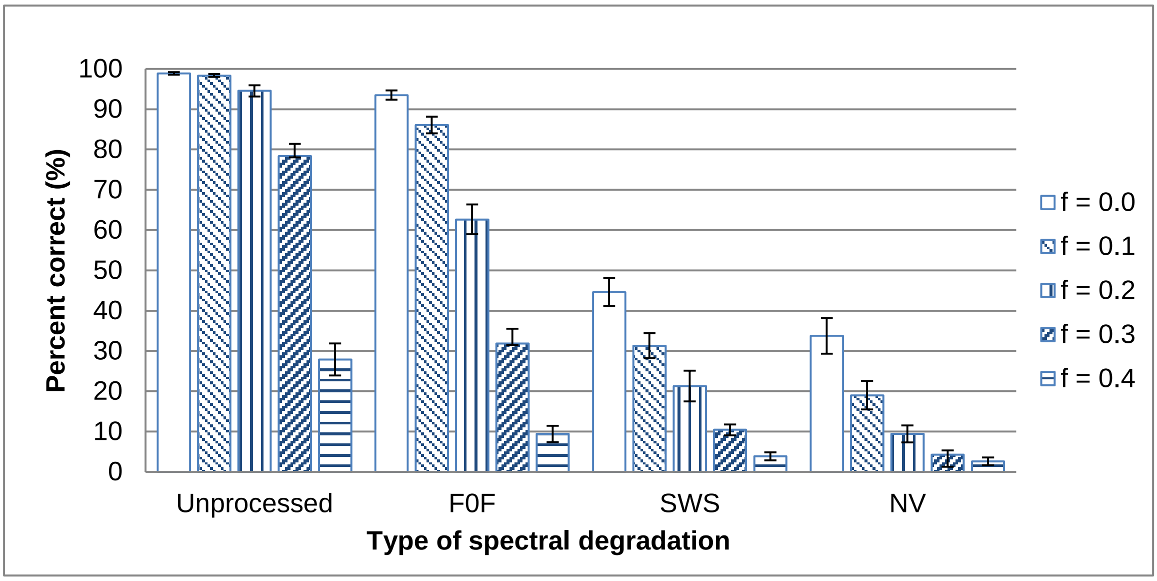

Figure 1. Mean sentence intelligibility score in percentage of (a) unprocessed *V*-only, (b) F_0F-V, (c) *SWS-V*, and (d) *NV-V* sentences preserving 100%, 80%, 60%, 40% and 20% of vowel centers (f = 0.0, 0.1, 0.2, 0.3 and 0.4 respectively). The error bars denote standard error of the mean.

3.1. Main effect: Spectral degradation

Multiple paired comparisons corrected using Bonferroni adjustment [significance level set at p < .008 (α = .05)] were conducted to investigate the main effect of spectral degradation with the same proportion factor f, and results revealed that the intelligibility scores of F_0F-V [F(1, 19) = 571.38, r = .98], *SWS-V* [F(1, 19) = 1587.24, r = .99] and *NV-V* [F(1, 19) = 1234.37, r = .99] sentences are all significantly lower than that of unprocessed *V*-only sentences, which served as control. This indicates that spectral degradation of *F_0F*, *SWS*, and *NV* would significantly lower intelligibility scores of *V*-only sentences.

3.1.1. Comparison of intelligibility scores of *V*-only and *F_0F-V* sentences. Paired-sample t-tests were carried out to compare the intelligibility scores of *V*-only and *F_0F-V* sentences at the same value of proportion factor *f*. The intelligibility scores of *F_0F-V* sentences are significantly lower than those of *V*-only sentences at all values of *f* ($p < .008$; $\alpha = .05$), as shown in *Table 1*. For instance, at $f = 0.0$ where full vowel segments are preserved, the mean intelligibility score of unprocessed condition is 98.9% (*SE* = 0.3), while that of *F_0F-V* condition is 93.5% (*SE* = 1.2). In terms of acoustic analysis, information about *F_0* contour is absent in *F_0F-V* sentences. Therefore, the result indicates that *F_0* contour is an acoustic cue significantly contributing to the intelligibility of *V*-only sentence.

3.1.2. Comparison of intelligibility scores of *F_0F-V* and *SWS-V*; and *F_0F-V* and *NV-V* sentences. Paired-sample t-tests were carried out to compare the intelligibility scores of *F_0F-V* and *SWS-V*, and *F_0F-V* and *NV-V* sentences with the same proportion factor *f*. The results are shown in *Table 1*. Intelligibility scores of *SWS-V* sentences are significantly lower than those of *F_0F-V* sentences at $f = 0.0$, 0.1, 0.2 and 0.3 ($p < .008$; $\alpha = .05$). Similarly, intelligibility scores of *NV-V* sentences are significantly lower than those of *F_0F-V* sentences at all values of proportion factor *f* ($p < .008$; $\alpha = .05$). For instance, at $f = 0.0$ where full vowel segments are preserved, the mean intelligibility scores of *F_0F-V* condition is 93.5%

(*SE* = 1.2), while those of *SWS-V* and *NV-V* conditions are 44.6% (*SE* = 3.5) and 33.8% (*SE* = 4.4) respectively. In terms of acoustic analysis, information about harmonics is absent in *SWS-V* and *NV-V* sentences when compared to F_0F-V sentences. This implies that harmonic cues significantly contribute to *V*-only sentence intelligibility.

3.1.3. Comparison of intelligibility scores of *SWS-V* and *NV-V* sentences.

Similarly, paired-sample t-tests were carried out to compare the intelligibility scores of *SWS-V* and *NV-V* sentences with the same proportion factor *f*. The intelligibility scores of *NV-V* sentences are significantly lower than that of *SWS-V* sentences at all values of *f*, as shown in *Table 1* ($p < .008$; $\alpha = .05$). For instance, at $f = 0.0$ where full vowel segments are preserved, the mean intelligibility scores of *SWS-V* condition is 44.6% (*SE* = 3.5), while that of *NV-V* condition is 33.8% (*SE* = 4.4). In terms of acoustic analysis, information about harmonics is absent in both *SWS-V* and *NV-V* sentences. However, more information about the first three formants is preserved in *SWS-V* sentences but none is preserved in *NV-V* sentences. The result seemly indicates that the first three formants are also acoustic cues contributing to *V*-only sentence intelligibility when harmonic structure is discarded.

3.1.4. Extent of contribution of F_0 contour and harmonic structure to *V*-only sentence intelligibility. In order to compare the extent of contribution of F_0 contour and harmonic structure to *V*-only sentence intelligibility, the differences in intelligibility scores in the above comparisons were investigated. Note that only results at $f = 0.0$, 0.1 and 0.2 were taken into account because of the high intelligibility in unprocessed *V*-only sentences (mean intelligibility score more than 90%). It was noted that the difference in intelligibility scores between unprocessed *V*-only and *F_0F-V* sentences [*f* = 0.0: (*M* = 15.17, *SE* = 2.47); *f* = 0.1: (*M* = 25.04, *SE* = 2.87); *f* = 0.2: (*M* = 41.67, *SE* = 4.03)] were much less than that between *F_0F-V* and *SWS-V* sentences [*f* = 0.0: (*M* = 56.73, *SE* = 4.46); *f* = 0.1: (*M* = 57.50, *SE* = 3.44); *f* = 0.2: (*M* = 42.88, *SE* = 3.81)] or *F_0F-V* and *NV-V* sentences [*f* = 0.0: (*M* = 68.42, *SE* = 4.80); *f* = 0.1: (*M* = 73.91, *SE* = 4.84); *f* = 0.2: (*M* = 60.16, *SE* = 3.95)] at *f* = 0.0, 0.1 and 0.2. This indicates that further extracting harmonic cues from *F_0F-V* sentences lowers sentence intelligibility to a greater extent than flattening *F_0* in unprocessed *V*-only sentences.

Table 1. Results of paired-sample t-tests comparing intelligibility scores of (a) unprocessed *V*-only vs. *F_0F-V*, (b) *F_0F-V* vs. *SWS-V*, (c) *F_0F-V* vs. *NV-V*, and (d) *SWS-V* vs. *NV-V* sentences.

Vowel duration[a]	**Mean of score difference**	**Standard error of mean**	**t-value**	**Significance value (2-tailed)**[b]	**Effect size, *r***
(a) Unprocessed *V*-only vs. *F_0F-V*					
f = 0.0	15.168	2.466	6.151	****p* < .001	.816
f = 0.1	25.037	2.871	8.719	****p* < .001	.894
f = 0.2	41.667	4.031	10.337	****p* < .001	.921
f = 0.3	48.226	4.372	11.031	****p* < .001	.930
f = 0.4	24.669	4.879	5.056	****p* < .001	.757
(b) *F_0F-V* vs. *SWS-V*					
f = 0.0	56.725	4.460	12.718	****p* < .001	.946
f = 0.1	57.503	3.435	16.743	****p* < .001	.968
f = 0.2	42.879	3.811	11.252	****p* < .001	.932

f = 0.3	25.435	4.014	6.337	****p* < .001	.824
f = 0.4	10.528	3.711	2.837	**p* = .011	.545
(c) *F_0F-V* vs. *NV-V*					
f = 0.0	68.418	4.800	14.255	****p* < .001	.956
f = 0.1	73.912	4.843	15.262	****p* < .001	.962
f = 0.2	60.164	3.953	15.221	****p* < .001	.961
f = 0.3	38.547	3.782	10.193	****p* < .001	.919
f = 0.4	15.704	2.851	5.509	****p* < .001	.784
(d) *SWS-V* vs. *NV-V*					
f = 0.0	11.693	3.401	3.438	***p* = .003	.619
f = 0.1	16.408	5.175	3.171	***p* = .005	.588
f = 0.2	17.284	3.167	5.457	****p* < .001	.781
f = 0.3	13.112	2.079	6.306	****p* < .001	.823
f = 0.4	5.176	2.344	2.208	***p* = .004	.452

Notes: [a] *f* = 0.0, 0.1, 0.2, 0.3 and 0.4 represents preservation of 100%, 80%, 60%, 40% and 20% of vowel centers respectively. [b] **p* < .05. ***p* < .01. ****p* < .001.

3.2. Main effect: Vowel duration

Statistical analysis indicated a significant main effect of vowel duration. In order to compare the effect of vowel duration on *V*-only sentence intelligibility with and without spectral degradation, paired-sample t-tests were conducted to compare the intelligibility scores under the same condition of spectral degradation. The Bonferroni-corrected statistical significance level was set at $p < .005$ ($\alpha = .05$). The results comparing the difference of intelligibility scores at $f = 0.0$ and 0.1 are shown in *Table 2*. It was observed that the effect of reducing vowel duration from 100% to 80%, i.e. $f = 0.0$ to $f = 0.1$, is insignificant in unprocessed *V*-only sentences [$t(19) = 1.19$, $p = .25$, $r = .26$]. However, the same reduction in vowel duration from 100% to 80% significantly reduces intelligibility in

F_0F-V sentences [*t*(19) = 5.18, *p* < .001, *r* = .77], *SWS-V* sentences [*t*(19) = 3.27, *p* = .004, *r* = .60], and *NV-V* sentences [*t*(19) = 4.59, *p* < . 001, *r* = .73]. Thus, the effect of reducing vowel duration from 100% to 80% contributes to a greater extent in lowering *V*-only sentence intelligibility under conditions of spectral degradation than that without spectral degradation. This may imply that the contribution of vowel duration to *V*-only sentence intelligibility is greater under conditions of spectral degradation. Note that *Table 2* only compares intelligibility scores of the pair *f* = 0.0 and 0.1. Not surprisingly, further reduction in vowel onset and offset durations (more than 10% from both vowel onset and offset) significantly reduces the intelligibility scores for all signal processing conditions.

Table 2. Results of paired-sample t-tests comparing intelligibility scores when 100% (*f* = 0.0) and 80% (*f* = 0.1) of vowel centers are preserved under different conditions of spectral degradation.

Spectral degradation	**Mean of score difference**	**Standard error of mean**	**t-value**	**Significance value (2-tailed)**[a]	**Effect size, *r***
Unprocessed *V*-only	2.688	2.265	1.186	*p* = .250	.263
F_0F-V	12.557	2.423	5.183	****p* < .001	.765
SWS-V	13.336	4.075	3.273	***p* = .004	.600
NV-V	18.051	3.933	4.589	****p* < .001	.725

Note: [a] ***p* < .01. ****p* < .001.

4. Discussion

The present study arises from previous research findings that supported a 3:1 *V* over *C* advantage in Mandarin sentence intelligibility (Chen et al., 2013). We investigate the effect of different acoustic cues (F_0 contour, harmonics and formants) and preservation of different durations of vowel centers on Mandarin *V*-only sentence intelligibility using the noise-replacement paradigm. The results suggest F_0 contour, harmonic structure and the first three formants to be acoustic cues contributing to *V*-only sentence intelligibility to different extents; thus they might account for the *V* advantage found. Vowel duration is also a factor significantly contributing to intelligibility under conditions of spectral degradation.

4.1. Effect of different acoustic cues

The results from previous findings with full consonant and vowel durations showed that high intelligibility could be maintained under conditions of spectral degradations of F_0F, *SWS* and *NV* (Chen & Loizou, 2011; Chen et al., 2014; Feng et al., 2012). Native listeners could comprehend spectrally degraded sentences through a top-down approach by using their syntactic and semantic knowledge of the language. Despite that, the present results showed that spectral degradations of F_0F, *SWS* and *NV* significantly reduce *V*-only sentence intelligibility, and the extent in reduction depends on the processing strategy used. This implies that segmental interruption (i.e., only preserving *V* in this study) does have a significant effect on the comprehensibility of spectrally degraded sentences. In *V*-only sentences, acoustic cues of *C*, which convey information about manners of starting and stopping of most Mandarin words, are discarded. We therefore presume that listener's ability in using semantic and syntactic features of language to compensate for the loss of acoustic cues under conditions of spectral degradation in *V*-only sentences is lowered without the information carried by *C*. Thus, the loss of acoustic information, such as F_0 contour, harmonics and formants, which limits perceptual cues that are available for sentence comprehension, resulted in significant decrease in intelligibility of *V*-only sentences.

4.1.1 Effect of fundamental frequency contour. F_0 contour is one of the dominant cues to tone recognition (e.g. Howie, 1976; Whalen & Xu, 1992). However, the present study suggests that extracting F_0 contour cue has a minimal effect in reducing *V*-only sentence intelligibility. Although F_0F manipulation significantly decreases intelligibility compared to that in unprocessed *V*-only sentences, high intelligibility score (93.5%) is still resulted in F_0F-V sentences at $f = 0.0$. This is consistent with previous findings that lexical tone are relatively redundant cues for intelligibility in quiet conditions as acoustic distinction of tones is compromised in natural speech, and other cues could compensate for the distorted lexical tone contours (Chen et al., 2014; Feng et al., 2012; Xu, 2006). Therefore, dynamic cues of F_0 only account minimally for the *V* advantage in sentence comprehension in Mandarin, as in English (Fogerty & Humes, 2012).

4.1.2. Effect of harmonic structure. In both *SWS* and *NV* processing, almost all of the harmonic cues are discarded when comparing *SWS*-*V* or *NV*-*V* sentences to F_0F-V

sentences. The results showed significant decrease in intelligibility when harmonic cues are discarded. This implies that harmonics cues may account for the *V* advantage found in sentence intelligibility. In sentence comprehension, context-dependent top-down processing is an important consideration in addition to acoustic cues, as in English. We hypothesize that harmonic structure might provide cues adequate for using top-down processing to compensate for the loss in F_0 contour information. Further study in effect of harmonics on word recognition would help to verify this hypothesis, as top-down processing could not be used in word recognition.

4.1.3. Effect of formant. Most of the harmonic cues are discarded in *SWS* and *NV* processing, but *SWS* processing still preserves some information about the first three formants, whereas *NV* processing does not. As shown by the significant decrease in intelligibility in *NV*-*V* sentences compared to *SWS*-*V* sentences, extracting information about first three formants seems to decrease *V*-only sentence intelligibility significantly when harmonic structure is discarded. However, even if present, its effect is minimal, given that the intelligibility scores only reduce slightly. For instance, the scores only reduce from 44.6% in *SWS*-*V* sentences to 33.8% in *NV*-*V* sentences at $f = 0.0$. The contribution of formant structure to *V*-only sentence intelligibility might be accounted by its effect on top-down processing of sentences. Information about formants in vowel segments including formant transition is captured in the temporal fine structure (Rosen, 1992). Temporal fine structure was found to enhance the advantage of envelope cues for the context-dependent advantage of *V* in English sentences (Fogerty & Humes, 2012). Likewise, the present results about formant contribution might demonstrate its positive effect on top-down processing in Mandarin *V*-only sentences.

4.1.4. Comparison of effects of F_0 contour and harmonic structure. At *f* = 0.0, 0.1 and 0.2, difference in intelligibility scores between unprocessed *V*-only and F_0F-*V* sentences are lower than that between F_0F-*V* and *SWS*-*V* or *NV*-*V* sentences. For instance, intelligibility score is only lowered from 98.9% to 93.5% when F_0 was flattened in *V*-only sentences at *f* = 0.0. However, when harmonic cues were further discarded as in *SWS*-*V* sentences, intelligibility score greatly reduces from 93.5% in F_0F-*V* sentences to 44.6% in *SWS*-*V* sentences. Similarly, a relatively great reduction in intelligibility was observed in *NV*-*V* sentences (33.8%) compared to F_0F-*V* sentences (44.6%). The results seem to suggest that harmonic structure has a greater extent of effect on *V*-only sentence intelligibility relative to F_0 contour. Therefore, we presume that harmonic structures carried by the vowel portion account for, to a greater extent, additional *V* advantage in sentence comprehension.

4.2. Effect of vowel center duration

Vowel duration is also one of the contributors to tone recognition (Whalen & Xu, 1992). The results showed that reducing vowel duration from 100% to 80% does not significantly reduce sentence intelligibility without spectral degradations (i.e. in unprocessed condition).

This is also consistent with previous studies that lexical tone is not an important cue for intelligibility under quiet conditions. Not surprisingly, further reduction in vowel duration (from preserving 80% to 20% of vowel centers) would significantly decrease intelligibility. We hypothesize that this further reduction in vowel duration would remove durations of other acoustic cues, such as harmonic structure in vowel portion, thus creating a severely interrupted signal. Therefore intelligibility is significantly reduced.

On the other hand, unlike unprocessed *V*-only sentences, reduction of vowel duration from 100% to 80% in F_0F-*V, SWS-V* and *NV-V* sentences significantly decreases intelligibility. This shows a significant impact of segment duration on *V*-only sentence intelligibility in the presence of spectral degradation. Listeners could use perceptual cues to compensate for the loss of vowel edges when there is no spectral degradation. Nevertheless, we hypothesize that this compensatory ability in understanding *V*-only sentences with shortened duration is negatively affected by conditions of spectral degradation.

4.3. Implication

The present study indicates the importance of preserving vowel duration and harmonic structure to sentence intelligibility under conditions of spectral degradation. As different degrees of spectral degradation are often experienced by hearing impaired patients, future design of novel speech processing strategies, such as those for assistive hearing devices, should consider techniques for preserving as much vowel portion as possible in order to maximize speech comprehensibility for segmentally interrupted sentences. Also, further research and development in speech transmission and amplification technologies should emphasize more on preservation of harmonic structures than of other acoustic cues in the vowel portion so as to maximize speech intelligibility.

4.4. Limitation of the present study

Although effect size has been taken into consideration, the present study has a relatively small sample size. Also, the individuals participating in the study differed in their cities they were staying in before coming to Hong Kong, which might affect their listening experience and expectation in Mandarin. This might influence their perceptual use of acoustic cues in sentence comprehension. Individual variations might affect generalizability of the findings to the general Mandarin-speaking population.

In addition to participant variations, identification of acoustic cues is complicated that individual cues cannot be extracted solely. For example, amplitude contour was found to be correlated to F_0 contour (Whalen & Xu, 1992). In our study, although conclusions for F_0 contour, harmonic structure and formant frequencies have been drawn by manipulating these cues broadly, interactions among them and with other acoustic cues could not be overlooked.

4.5. Future study

The present study was carried out under quiet listening conditions. Some studies found that acoustic cues such as F_0 contour were important to speech intelligibility under condition with noise (e.g. Chen et al., 2013; Patel, Xu, & Wang, 2010). In order to generalize the findings to facilitate the design of hearing assistive devices used in daily situations, it is suggested to investigate the effects in noise. Also, further study could include the elderly, as acoustic cues may be processed differently due to cognitive decline (Humes, 2007). Hearing impaired individuals, who usually experience different degrees and types of spectral degradations, could also be included in future studies.

In addition, the test stimuli were processed from sentence materials in *MHINT*, which represents a simple conversational speech easily understood by native Mandarin listeners with a variety of educational backgrounds. The findings in the present study thus represent the phenomenon in comprehending high-probability sentences. In low probability sentences, we suppose listener's ability in using a top-down approach for comprehension is lowered because of unfamiliar semantic cues, so tone information might become more crucial for enhancing intelligibility. Therefore, future studies can investigate the effect of F_0 contour and harmonic cues on low-probability sentences.

5. Conclusion

The present study investigated the contribution of F_0 contour, harmonic structure and formant to Mandarin *V*-only sentence intelligibility. It was shown that removing each of these acoustic cues would significantly decrease intelligibility of *V*-only sentences to different extents. Consistent with previous findings in languages other than Mandarin, removing harmonic cues, either by sine-wave synthesis or noise-vocoding processing, would decrease intelligibility significantly, whereas flattening fundamental frequency cues would have a minimal effect on reducing intelligibility. Thus harmonic cues might account for the vowel advantage found in Mandarin sentence comprehension. The present work also showed that spectral degradation would lower listener’s ability to compensate for the loss of vowel edges to comprehend vowel-only sentences.

6. Acknowledgement

The study was supported by Faculty Research Fund, Faculty of Education, The University of Hong Kong and Seed Funding for Basic Research, The University of Hong Kong. The author would like to acknowledge all participants of this study, and express thankfulness to the author’s supervisor, Dr. Felix Chen, for his kind and supportive guidance.

7. References

Allen, J. B. (1994). How do humans process and recognize speech? *IEEE Transactions on Speech and Audio Processing, 2,* 567-577. doi: 10.1109/89.326615

American National Standards Institute (1996). *American national standard specifications for audiometers*. New York: Acoustical Society of America.

Chen, F., & Loizou, P. (2011). Predicting the intelligibility of vocoded and wideband Mandarin Chinese. *Journal of the Acoustical Society of America, 129,* 3281-3290. doi: 10.1121/1.3570957

Chen, F., Wong, L. N., & Hu, Y. (2014). Effects of lexical tone contour on Mandarin sentence intelligibility. *Journal of Speech, Language, and Hearing Research, 57,* 338-345. doi: 10.1044/1092-4388(2013/12-0324)

Chen, F., Wong, L.N., & Wong, Y. W. (2013). Assessing the perceptual contributions of vowels and consonants to Mandarin sentence intelligibility. *Journal of the Acoustical Society of America, 134,* EL178-EL184. doi: 10.1121/1.4812820

Feng, Y.M., Xu, L., Zhou, N., Yang, G., & Yin, S. K. (2012). Sine-wave speech recognition in a tonal language. *Journal of the Acoustical Society of America, 131,* EL133–EL138. doi: 10.1121/1.3670594

Fogerty, D. (2012). Importance of sentence-level and phoneme-level envelope modulations during vowels in interrupted speech. *Proceedings of Meetings on Acoustics, 18,* 06002.

doi: 10.1121/1.4772391

Fogerty, D., & Humes, L.E. (2012). The role of vowel and consonant fundamental frequency, envelope, and temporal fine structure cues to the intelligibility of words and sentences. *Journal of the Acoustical Society of America, 131,* 1490-1501. doi: 10.1121/1.3676696

Fogerty, D., & Kewley-Port, D. (2009). Perceptual contributions of the consonant-vowel boundary to sentence intelligibility. *Journal of the Acoustical Society of America, 126,* 847-857. doi: 10.1121/1.3159302

Fu, Q. J., Zeng, F. G., Shannon, R. V., & Soli, S. D. (1998). Importance of tonal envelope cues in Chinese speech recognition. *Journal of the Acoustical Society of America, 104,* 505-510. doi: 10.1121/1.423251

Hill, F. J., McRae, L. P., & McClellan, R. P. (1968). Speech recognition as a function of channel capacity in a discrete set of channels. *Journal of the Acoustical Society of America, 44,* 13-18. doi: 10.1121/1.1911047

Howie, J. M. (1976). *Acoustical studies of Mandarin vowels and tones.* Cambridge: Cambridge University Press.

Humes, L. E. (2007). The contributions of audibility and cognitive factors to the benefit provided by amplified speech to older adults. *Journal of the American Academy of Audiology, 18,* 590-603. doi: 10.3766/jaaa.18.7.6

Kewley-Port, D., Burkle, T. Z., & Lee, J. H. (2007). Contribution of consonant versus vowel information to sentence intelligibility for young normal-hearing and elderly hearing-impaired listeners. *Journal of the Acoustical Society of America, 122,* 2365-2375. doi: 10.1121/1.2773986

Laures, J. S., & Weismer, G. (1999). The effects of a flattened fundamental frequency on intelligibility at the sentence level. *Journal of Speech, Language, and Hearing Research, 42,* 1148-1156. doi: 10.1044/jslhr.4205.1148

Loizou, P. C., Dorman, M., & Tu, Z. (1999). On the number of channels needed to understand speech. *Journal of the Acoustical Society of America, 106,* 2097-2103. doi: 10.1121/1.427954

Patel, A. D., Xu, Y., & Wang, B. (2010). The role of F0 variation in the intelligibility of Mandarin sentences. *Speech Prosody 2010 100890.* Retrieved from http://20.210-193-52.unknown.qala.com.sg/archive/sp2010/papers/sp10_890.pdf

Remez, R. E., Rubin, P. E., Berns, S. M., Pardo, J. S., & Lang, J. M. (1994). On the perceptual organization of speech. *Psychological Review, 101*, 129-156. doi: 10.1037/0033-295X.101.1.129

Rosen, S. (1992). Temporal information in speech: acoustic, auditory, and linguistic aspects. *Philosophical Transactions of the Royal Society B, 336,* 367-373. doi: 10.1098/rstb.1992.0070

Studebaker, G. A. (1985). A ‘rationalized’ arcsine transform. *Journal of Speech and Hearing*

Research, 28, 455–462. doi: 10.1044/jshr.2803.455

Waibel, A. (1987). Prosodic knowledge sources for word hypothesization in a continuous speech recognition system. *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP '87, 12,* 856-859. doi: 10.1109/ICASSP.1987.1169848

Whalen, D. H., & Xu, Y. (1992). Information for Mandarin tones in the amplitude contour and in brief segments. *Phonetica, 49,* 25-47. doi: 10.1159/000261901

Wong, L. L., Soli, S. D., Liu, S., Han, N., & Huang, M. W. (2007). Development of the Mandarin Hearing in Noise Test (MHINT). *Ear and Hearing, 28,* 70S-74S. doi: 10.1097/AUD.0b013e31803154d0

Xu, Y. (2006). Tone in connected discourse. In K. Brown (Ed.), *Encyclopedia of Language and Linguistics (2nd Ed.).* Retrieved from http://www.phon.ucl.ac.uk/home/yi/yispapers/Xu_ELL_author_version.pdf