INTERSPEECH 2021
30 August – 3 September, 2021, Brno, Czechia

Perceptual Contributions of Vowels and Consonant-vowel Transitions in Understanding Time-compressed Mandarin Sentences

Changjie Pan [1], *Feng Yang* [2], *Fei Chen* [1]

[1] Department of Electrical and Electronic Engineering, Southern University of Science and Technology, Shenzhen, China
[2] Department of Stomatology, Shenzhen Second People's Hospital, The First Affiliated Hospital of Shenzhen University, Shenzhen, China
fchen@sustech.edu.cn

Abstract

Many early studies reported the importance of vowels and vowel-consonant transitions to speech intelligibility. The present work assessed their perceptual impacts to the understanding of time-compressed sentences, which could be used to measure the temporal acuity during speech understanding. Mandarin sentences were edited to selectively preserve vowel centers or vowel-consonant transitional segments, and compress the rest regions with equipment time compression rates (TCRs) up to 3, including conditions only preserving vowel centers or vowel-consonant transitions. The processed stimuli were presented to normal-hearing listeners to recognize. Results showed that, consistent with the segmental contributions in understanding uncompressed speech, the vowel-only time-compressed stimuli were highly intelligible (i.e., intelligibility score >85%) at a TCR around 3, and vowel-consonant transitions carried important intelligibility information in understanding time-compressed sentences. The time-compression conditions in the present work provided higher intelligibility scores than their counterparties in understanding the PSOLA-processed time-compressed sentences with TCRs around 3. The findings in this work suggested that the design of time compression processing could be guided towards selectively preserving perceptually important speech segments (e.g., vowels) in the future.

Index Terms: time compression, speech intelligibility, pitch synchronous overlap and add (PSOLA)

1. Introduction

The segmental (e.g., vowels) contributions to speech intelligibility have been long studied, and consistent findings have been reported with the usage of a noise-replacement paradigm [e.g., 1-4]. In implementing the noise-replacement paradigm, the original speech signal is first separated into various segments (e.g., vowels and consonants, or Vs and Cs) with available acoustic landmark information. The segments of interest are preserved, while the rest segments are replaced (not compressed or discarded) with noise. Note that the duration of the noise-replaced stimulus is the same as that of the original speech. Early work consistently showed that vowels carry more perceptual information than consonants. Vowel-only sentences (i.e., consonants are replaced with noise) are more intelligible than consonant-only sentences (i.e., replacing vowels with noise), yielding an intelligibility ratio of 2 in English [1-3] and a much higher ratio in Mandarin Chinese [4]. In addition, consonant-vowel (C-V) transitions were found to carry important perceptual information, and adding a small amount of C-V transitions to consonant-only sentences may significantly improve speech understanding [3-4]. Recently, the perceptual importance of selected speech segments has been extended to many other studies. For instance, Chen and Chen showed that the combined advantage in electric-and-acoustic hearing could be achieved when only vowel onsets were provided [5].

The noise-replacement processing preserves the duration of the replaced segments (with noise), which motivated us to investigate the importance of those noise-replaced segments. In other words, as those noise-replaced segments do not contain acoustic waveforms, how would they affect speech understanding if the noise-replaced segments are compressed or discarded (yielding a time-compressed speech)? Time-compression of speech has been commonly used to study speech perception and age-related speech processing deficits (e.g., temporal acuity), particularly in those hearing-impaired listeners [e.g., 6-10]. Grose et al. studied the modulation masking release under various time-compression processing, and found that speech perception thresholds increased with increasing time-compression in two maskers (a steady speech-shaped noise masker and a modulated masker), but more markedly in the modulated masker [7]. Banai and Lavner studied the effects of exposure and training on the perception of time-compressed speech in native versus nonnative listeners, and found that learning profiles differed between native and nonnative listeners and exposure had a weaker effect in nonnative than in native listeners [8]. Meng et al. recently compared the perception of time-compressed speech between normal-hearing (NH) and cochlear implant (CI) listeners, and their results showed that the time-compression thresholds were around 16.7 syllables/sec for normal listeners, but rarely faster than 10.0 syllables/sec for implanted listeners [9]. Li et al. compared speech perception obtained with different time compression rates (TCRs) in teenagers that did or did not use personal listening devices (PLDs), and found that the fast-speed speech recognition in noise decreased significantly in PLD users compared with that in non-PLD users selected by extreme entertainment exposure [10].

The quantification of time-compression is usually expressed in terms of the proportion of the time-waveform content that is excised. For example, 33%-time-compression implies that one-third of the original time waveform has been removed, yielding an equivalent TCR (i.e., the ratio between the durations of the original uncompressed and time-compressed speech waveforms) of 1.5. Time compression manipulations typically remove redundant (or perceptually

http://dx.doi.org/10.21437/Interspeech.2021-58

less-important) speech segments aiming to increase speech perception rate but not markedly generate speech distortions causing detrimental impact to speech perception. Versfeld and Dreschler reported that young NH listeners could understand 50% of the sentences at a speaking rate of 12.5 syllables/sec when the speech materials were time-compressed by pitch synchronous overlap and add (PSOLA) [11, 12]. In [13], speech intelligibility tests were taken at the normal, slow or fast speaking rates. The speech materials at slow or fast speaking rates were processed by PSOLA to generate sentence stimuli at either half-speed (slow) or double-speed (fast). Their results indicated that slowing down the speaking rate had no significant impact on the speech recognition performances, but doubling the speaking rate was detrimental to the performances. The speech recognition scores were less than 10% for CI listeners and were around 80% for NH listeners at double-speed speaking rate. Johnson et al. recently found that time compression decreased intelligibility for all talkers, but the effect was significantly greater for speech of female talkers [14].

The aim of the present work was to extend early work on the perceptual importance of vowels and C-V transitions in understanding time-compressed sentences. Given the relative perceptual importance of vowels and vowel-consonant transitions, we hypothesized that when the time-compression processing was only applied to the consonant and pause segments, or vowel segments (e.g., vowel centers) were preserved, the time-compressed stimuli were still highly intelligible.

2. Methods

2.1. Subjects and materials

This experiment involved eight NH listeners (6 males and 2 females; pure-tone thresholds better than 20 dB hearing level at octave frequencies of 125–8000 Hz in both ears). All subjects were native speakers of Mandarin Chinese and were paid for their participation. The experimental procedure involving human subjects was approved by the Institution’s Ethical Review Board of Southern University of Science and Technology.

The speech material comprised sentences extracted from the Mandarin Hearing in Noise Test (MHINT) database [15]. The MHINT corpus includes 24 lists, each with 10 sentences and 10 keywords per sentence. All sentences were spoken by a male native Mandarin Chinese speaker with a fundamental frequency (F0) of 75–180 Hz and recorded at a sampling rate of 16 kHz.

2.2. Signal processing

The original uncompressed sentences were first processed to generate segmentally manipulated stimuli as follows. The C-V boundaries (defined based on traditional segmental boundaries) were labeled manually by an experienced phonetician, and later verified by another experienced phonetician [4-5]. Note that all final nasal Cs were counted as parts of the preceding Vs. According to [16], Mandarin Chinese has 21 Cs ([p], [ph], [m], [f], [t], [th], [n], [l], [k], [kh], [x], [tɕ], [tɕh], [ɕ], [tʂ], [tʂh], [ʂ], [ʐ], [ts], [tsh], and [s]) and 35 Vs ([a], [o], [ɤ], [i], [u], [y], [aɪ], [eɪ], [ɑʊ], [oʊ], [ia], [iɛ], [iaʊ], [iəʊ], [ua], [uo], [uaɪ], [ueɪ], [yɛ], [an], [ən], [ɑŋ], [əŋ], [oŋ], [iɛn], [in], [iɑŋ], [iŋ], [iʊŋ], [uan], [uən], [uɑŋ], [ʊəŋ], [yɛn], and [yn]). The 35 Vs consist of 6 simple Vs, 13 complex Vs, and 16 compound nasal Vs [16]. V nuclei with final nasals were grouped as Vs, as in many earlier Mandarin studies [e.g., 16-18]. The target segments (e.g., Vs or Cs) were preserved, according to the acoustic landmarks identified. Then, the rest segments were replaced with a speech-shaped noise scaled to 16 dB below the level of the intact speech waveform [e.g., 3]. These two steps implemented the traditional noise-replacement paradigm. To generate the speech-shaped noise, a finite impulse response filter was designed based on the average spectrum of all MHINT sentences, and a white noise was filtered to have the same long-term average spectrum as the MHINT sentences. Finally, controlled with a time compression (TC) value, the noise-replaced segments were compressed in duration. Four TC values were used in this work to generate four time-compression conditions, including TC=100%, 66%, 33%, and 0%. TC value 100% denotes that the noise-replaced segments were not compressed, and TC value 33% denotes that the duration of the noise-replaced segments was compressed to one-third of the duration of the original (uncompressed) noise-replaced segments. Note that TC value 0% denotes that all noise-replaced segments were removed, and the generated stimuli only contained target segments.

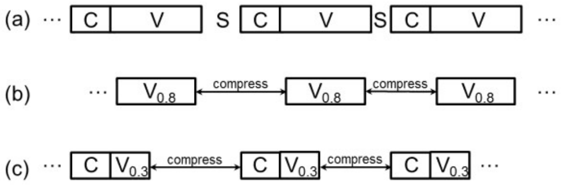

Figure 1: *Schema of the two time-compression conditions created. Panel (a) shows an uncompressed sentence with consonant (C), vowel (V) and silence (S). The time-compression condition in panel (b) preserves 80% vowel centers and compresses the rest noise-replaced segments. The time-compression condition in panel (c) preserves consonants plus 30% vowel onsets and compresses the rest noise-replaced segments.*

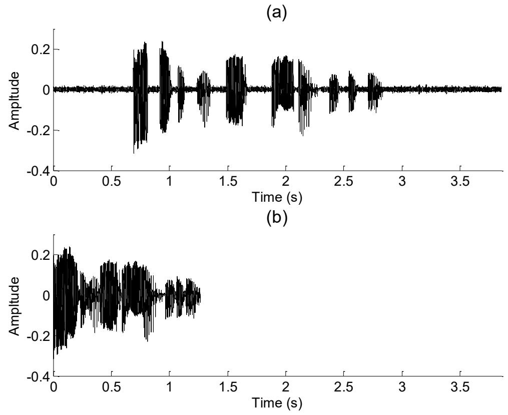

Figure 2: *Example waveforms of (a) an uncompressed noise-replaced (i.e., $V_{0.8}$ (TC=100%)) sentence, and (b) a time-compressed sentence with concatenated vowel centers.*

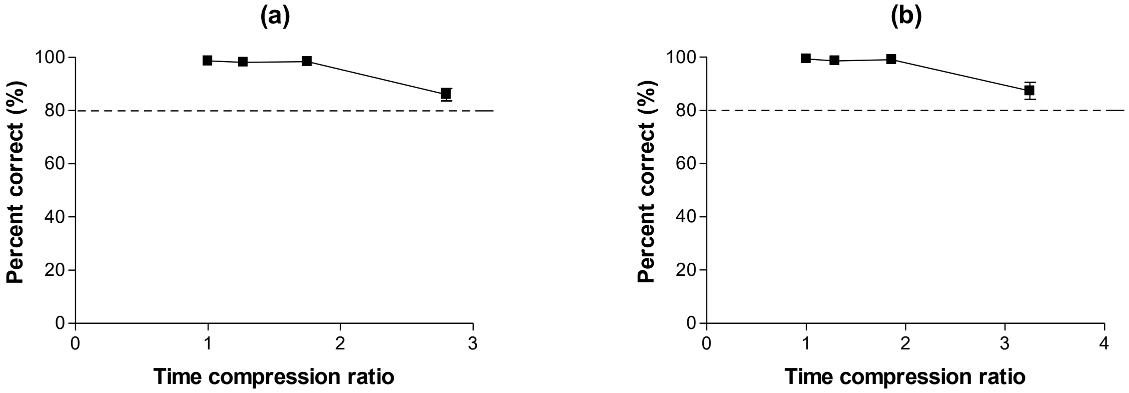

Figure 3: *Mean sentence recognition scores for all (a) $V_{0.8}$ and (b) $C+V_{0.3}$ conditions. The error bars denote ±1 standard error of the mean.*

Two types of segmentally manipulated stimuli were generated: 1) V center, and 2) C plus V onset (see Fig. 1). For the V-center condition, 80% of centered vowel segments were preserved, or 10% vowel onsets and 10% vowel offsets were not preserved. Hence, the V-center condition was also noted as the $V_{0.8}$ condition, whereas the subscript denotes the amount of vowels preserved from vowel center. The reason to generate the $V_{0.8}$ condition is two folds. First, due to the co-articulation effect, vowel onsets may contain some perceptional information on initial consonants in Mandarin words. Removing vowel onsets may largely (if not completely) diminish this co-articulation effect. Second, early work showed that the Mandarin sentences at the $V_{0.8}$ condition had an almost perfect performance of speech perception [4]. For the C+V-onset condition, all consonants and 30% vowel onsets were preserved, yielding the $C+V_{0.3}$ condition, whereas the subscript denotes the amount of vowels preserved from vowel onset. Early work also showed that NH listeners could nearly perfectly understand the Mandarin sentences processed with the $C+V_{0.3}$ condition [4].

Figure 1 illustrates the schematics of these two segmental conditions. Note that the $V_{0.8}$ conditions with TC=100%, 66%, 33%, and 0% equivalently compress the original uncompressed speech signal with TCRs of 1.0, 1.3, 1.8 and 2.8, respectively; while the $C+V_{0.3}$ conditions with TC=100%, 66%, 33%, and 0% equivalently compress the original uncompressed speech signal with TCRs of 1.0, 1.3, 1.9 and 3.3, respectively. Hence, removing all noise-replaced segments in the $V_{0.8}$ and $C+V_{0.3}$ conditions yields an equivalent TCR around 3. Figure 2 illustrates the waveforms of an uncompressed noise-replaced (i.e., $V_{0.8}$ (TC=100%)) sentence and a time-compressed stimulus at the $V_{0.8}$ (TC=0%) condition.

2.3. Procedure

The experiment was conducted in a sound-proof booth. Participants listened to the stimuli through circumaural headphones at a comfortable volume. Before the testing session, each participant attended a 10-min training session conducted with two lists of 10 MHINT sentences (i.e., the $V_{0.8}$ and $C+V_{0.3}$ conditions with TC=100%) not used in the testing session. The training session familiarized the participants with the testing procedure and test conditions. During the training session, the subjects were allowed to read transcripts of the training sentences when listening to them. In the testing session, the lists and order of the test conditions were randomized across participants, and the participants orally repeated as many words as they could recognize. Each subject participated in 8 test conditions (= 2 types of segmental conditions [$V_{0.8}$ and $C+V_{0.3}$] × 4 time-compression values [TC=100%, 66%, 33%, and 0%]). One list of 10 Mandarin sentences was used per test condition, and no sentence was repeated across conditions. Participants were allowed to listen to each stimulus a maximum of three times. During the testing session, a tester accompanied the participant and scored his/her responses simultaneously. The intelligibility score for each condition was computed as the ratio between the number of correctly recognized words and the total number of words contained in each MHINT list. The total testing time was less than 1 hour.

3. Results

Statistical significance was determined using the percent recognition score as the dependent variable, and TCR as the within-subject factor. Recognition scores were converted to rational arcsine units using the transform of Studebaker [19]. Figure 3(a) shows the mean sentence recognition results for the $V_{0.8}$ conditions. One-way repeated-measures analysis of variance (ANOVA) revealed a significant effect of TCR ($F_{3,21} = 27.769$, $p < 0.001$). Figure 3(b) shows the mean sentence recognition results for the $C+V_{0.3}$ conditions. One-way repeated-measures ANOVA revealed a significant effect of TCR ($F_{3,21} = 13.759$, $p < 0.001$).

4. Discussion and conclusions

The present work assessed the relative importance of vowels and vowel-consonant transitions in understanding time-compressed Mandarin sentences. Consistent with early findings, it was revealed that when only vowel centers or C-V transitional segments were preserved in the time-compression conditions, the Mandarin sentences were still quite intelligible. Specially, this work showed that when only vowel centers (i.e., 80% of centered vowel segments) were concatenated (yielding an equivalent TCR of 2.8) and all other segments were removed, the intelligibility score of the compressed sentences was above 85% (see Fig. 3(a)). Similarly, when only C-V transitional segments (i.e., consonants plus 30% vowel onsets) were preserved and other segments were removed, yielding an equivalent TCR of 3.3, the intelligibility score of the time-compressed sentences was round 87% (see Fig. 3 (b)). Versfeld and Dreschler reported an intelligibility score of 80% at a TCR of 3.2 when young NH listeners were presented with the PSOLA-processed time-compressed English sentences (see Fig. 1 in [11]). In [13], it was reported that the sentence recognition score for NH listeners was just over 80% at a TCR of 2 by adopting the PSOLA-based time-compression technique, while this work (see Fig. 3) shows that at a TCR of 2, the sentence recognition scores are over 90%. These comparisons show the better speech understanding performance of the time-compression methods used in this work than those in [11, 13], and indicate that time-compressed sentences could preserve high intelligibility if preserving selected segments (e.g., vowel centers and C-V transitions) known to have more importance to sentence understanding. Hence, the outcomes of this work provide insights to enhance the perception of time-compressed sentences by selectively compressing the perceptually less-important segments or preserving perceptually important segments.

Several factors may account for the high intelligibility of the time-compressed speech dominated with vowel centers or C-V transitional segments. First, vowels or C-V transitional segments carry important intelligibility information, as revealed in several work studying the perceptual importance of vowels and C-V transitional segments [1-5]. Particularly, Mandarin is a tonal language, and its tonal contour (characterized by F0 trajectory) carries important information for sentence understanding [4]. Second, the top-down mechanism may play an important role in understanding time-compressed sentences. Though a large portion of speech segments are removed in the time-compressed speech, listener may still use his/her language experience and knowledge, and contextual information to guess the meaning of the time-compressed sentences, which is further facilitated by the rich perceptual information contained in the preserved vowels and C-V transitions.

Early work studied methods to improve the understanding of time-compressed sentences. For instance, Gordon-Salant et al. investigated how selective time expansion processing affected the intelligibility of time-compressed sentences, and found that selective time expansion of consonants applied to rapid speech (i.e., 50% time-compression) could provide perceptual benefit to old listeners and listeners with hearing impairment [20]. The findings of this work also provided insights for the design of time compression processing. As shown in this work, when the non-vowel segments were cut and only vowel segments were preserved (i.e., the $V_{0.8}$ (TC=0%) condition), the intelligibility score could reach 85%. If the boundaries of various vowel segments are known, the intelligibility score could be further improved by inserting weak noise segment between adjacent vowel segments (i.e., generating the $V_{0.8}$ (TC=100%) condition), yielding an almost perfect understanding of vowel-only sentences [4]. Hence, due to the large intelligibility information carried in vowels, speech compression and coding could selectively focus on vowel segments.

Note that the present work was conducted with Mandarin Chinese, which is a tonal language with a high relative importance of vowels for speech perception [4]. It is unknown whether studies with English materials may have the similar findings, which warrants a further comparative study across languages [21]. Mandarin has more vowels and a longer vowel duration than English. In addition, Mandarin is characterized with its monosyllabic word structure, compared to the multi-syllabic word structure in English. These differences between Mandarin and English may potentially impact the relative importance of vowels in understanding time-compressed English sentences.

In conclusion, the present work studied the perceptual contributions of selected speech segments in understanding time-compressed Mandarin sentences. Two speech segments were preserved in the time-compressed sentences, including vowel centers and consonant-vowel transitions, and other segments were compressed. Results showed that the time-compressed stimuli with concatenated vowel centers or consonants plus C-V transitions could better preserve Mandarin sentence intelligibility, and provided a better speech understanding performance (i.e., intelligibility score >85%) under a TCR around 3. The design of time compression processing could be guided towards selectively preserving perceptually important segments (e.g., vowels) in the future.

5. Acknowledgements

This work was supported by Tencent Ethereal Audio Lab, Tencent Corporation.

6. References

[1] Cole, R., Yan, Y., Mak, B., Fanty, M., and Bailey, T., “The contribution of consonants versus vowels to word recognition in fluent speech,” in Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), pp. 853–856, 1996.

[2] Kewley-Port, D., Burkle, T. Z., and Lee, J. H., “Contribution of consonant versus vowel information to sentence intelligibility for young normal-hearing and elderly hearing-impaired listeners,” J. Acoust. Soc. Am. 122, 2365–2375, 2007.

[3] Fogerty, D., and Kewley-Port, D., “Perceptual contributions of the consonant-vowel boundary to sentence intelligibility,” J. Acoust. Soc. Am. 126, 847–857, 2009.

[4] Chen, F., Wong, L. L. N., and Wong, Y. W., “Assessing the perceptual contributions of vowels and consonants to Mandarin sentence intelligibility,” J. Acoust. Soc. Am. 134, EL178–EL184, 2013.

[5] Chen, F., and Chen, J., “Perceptual contributions of vowels and consonant-vowel transitions in simulated electric-acoustic hearing”, J. Acoust. Soc. Am. 145, EL197–202, 2019.

[6] Gordon-Salant, S., Zion, D. J., and Espy-Wilson, C., “Recognition of time-compressed speech does not predict recognition of natural fast-rate speech by older listeners,” J. Acoust. Soc. Am. 136, EL268–274, 2014.

[7] Grose, J. H., Griz, S., Pacífico, F. A., Advíncula, K. P., and Menezes, D. C., “Modulation masking release using the Brazilian-Portuguese HINT: Psychometric functions and the effect of speech time compression,” Int. J. Audiol. 54, 274–281, 2015.

[8] Banai, K., and Lavner, Y., “The effects of exposure and training on the perception of time-compressed speech in native versus nonnative listeners,” J. Acoust. Soc. Am. 140, 1686–1696, 2016.

[9] Meng, Q. L., Wang, X. R., Cai, Y. X., Kong, F. H., Buck, A. N., Yu, G. Z., Zheng, N. H., and Schnupp, J. W. H., “Time-compression thresholds for mandarin sentences in normal-hearing and cochlear implant listeners,” Hear. Res. 374, 58–68, 2019.

[10] Li, K. Y., Xia, L., Zheng, Z., Liu, W. Q., Yang, X. Y., Feng, Y. M., and Zhang, C. X., “A preliminary study on time-compressed speech recognition in noise among teenage students who use personal listening devices,” Int. J. Audiol. 58, 125–131, 2019.

[11] Versfeld, N. J., and Dreschler, W. A., “The relationship between the intelligibility of time-compressed speech and speech in noise

in young and elderly listeners," J. Acoust. Soc. Am. 111, 401–408, 2002.

[12] Moulines, E., and Laroche, J. "Non-parametric techniques for pitchscale and time-scale modification of speech," Speech Commun. 16, 175–205, 1995.

[13] Ji, C., Galvin, J. J., 3rd, Xu, A., and Fu, Q. J., "Effect of speaking rate on recognition of synthetic and natural speech by normal-hearing and cochlear implant listeners," Ear Hear. 34, 313–323, 2013.

[14] Johnson, E. M., Morgan, S. D., and Ferguson, S. H., "Does Time Compression Decrease Intelligibility for Female Talkers More Than for Male Talkers?" J. Speech Lang. Hear. Res. 63, 1083–1092, 2020.

[15] Wong, L. L. N., Soli, S. D., Liu, S., Han, N., and Huang, M. W., "Development of the Mandarin Hearing in Noise Test (MHINT)," Ear Hear. 28, 70S–74S, 2007.

[16] Yin, B., and Felley, M., "Chinese Romanization: Pronunciation and Orthography," Sinolingua, Beijing, China, 1990.

[17] Fu, Q. J., Zhu, M., and Wang, X. S., "Development and validation of the Mandarin speech perception test," J. Acoust. Soc. Am. 129, EL267–EL273, 2011.

[18] Chen, F., Wong, M. L. Y., Zhu, S. F., and Wong, L. L. N., "Relative contributions of vowels and consonants in recognizing isolated Mandarin words," J. Phonetics 52, 26–34, 2015.

[19] Studebaker, G. A. "A 'rationalized' arcsine transform," J. Speech Hear. Res. 28, 455–462, 1985.

[20] Gordon-Salant, S., Fitzgibbons, P. J, and Friedman, S. A., "Recognition of time-compressed and natural speech with selective temporal enhancements by young and elderly listeners," J. Speech Lang. Hear. Res. 50, 1181–1190, 2007.

[21] Fogerty, D., and Chen, F. "Vowel spectral contributions to English and Mandarin sentence intelligibility," in Proceedings of 15th Annual Conference of the International Speech Communication Association (InterSpeech), pp. 499–503, 2014.