Effects of Temporal Modulation Coding Characteristics in the Auditory Midbrain On Speech Recognition

Mengna Sun, Tianhao Li, Yanji Jiang
Liaoning Technical University, Liaoning, China
E-mail:mengnasun_amy@qq.com Tel: 183 4299 8048
E-mail: tianhaol@gmail.com Tel:159 4296 6121

***Abstract*——Auditory midbrain neurons not only tune to frequencies but also tune to temporal modulations, coding temporal envelope modulation with rate or phase-locking responses. It is unclear whether and how these coding characteristics influence speech recognition. Based on our earlier observations on midbrain neurons' responses to complex harmonics with different phase relationships, we replaced the vowel segment in the sentence by complex harmonics with different starting phases. The complex harmonics kept similar spectral envelopes of original vowel segments and the individual harmonic's starting phase were manipulated to generate different temporal envelopes. Four kinds of starting phase (original phase, cosine phase, alternative phase and random phase) were utilized to re-synthesize the vowel segments. Twenty native speakers of Mandarin participated into the study. Within-subject design was used in this study and each subject listened to all test conditions in quiet. Sentence and word recognition scores were recorded to quantify the listeners' speech intelligibility. Results showed that recognition performances for processed sentences were significantly lower than that for original sentences, suggesting that degraded temporal fine structure in speech mitigates speech intelligibility, even in quiet. Recognition performance for the condition of random phase was generally lower than other conditions and recognition performance for the condition of cosine phase was the best, suggesting that rate and phase-locking responses to the temporal envelope in the auditory midbrain facilitate speech perception of the central neural system, even in quiet.**

I. INTRODUCTION

Speech sound consists of vowels and consonants across all languages. Mandarin Chinese resorts lexical tones locating at vowels to enlarge semantic range. Vowel sound is featured with rich harmonic structures and long duration. Consonant sound get short time duration and aperiodic spectrum. Auditory pathway as the media between sound input and listening understanding in the brain undertakes the role of sound processing. Auditory neurons produce nerve impulse synchronizing with the phase of periodical signal and tune to the frequency of signal. Phase-locking response has been verified as the messages encoding method in the auditory periphery system. Frequency following response is a kind of auditory evoked potential captured in scalp, which is mainly believed that the source of it is the phase-locking response in inferior colliculus (namely auditory midbrain). The spiking rates of nervous impulse in auditory midbrain neurons tune to frequency and temporal envelope, then convey into ascending central neural systems for further processing.

A number of researches have been executed to explore the speech intelligibility contribution of component. *Kewley-Port et al.* studied perception contribution of vowels versus consonants in English using noise replacement paradigm. Considering the difference of syllable structure and component distribution, *Chen et al.* executed similar experiment in Mandarin Chinese. Two experimental results presented that vowels made a more contribution for speech intelligibility than consonants. *Forgerty et al.* explored the perception contribution of consonant-vowel boundary (C-V boundary) using the same paradigm in English. It was found that speech intelligibility improved linearly with the increase of C-V

boundary proportion in the context of consonant and the intelligibility didn't decline saliently until 30% of vowels were replaced in the context of consonant. *Chen et al.* explored identical subject in Mandarin Chinese as *Forgerty et al.* in their experiment. Results got a nonlinear relationship between intelligibility and C+VP instead. The intelligibility contribution of C-V boundary indicated that a small portion of C-V boundary may improve intelligibility of C-only speech. Speech intelligibility can also be studied from the perspective of auditory information processing. *Laurel H. et al* (2013) simulated the population auditory nerve fibers responses and population auditory midbrain neurons responses for vowel space using valid computational models. Simulation results indicated that the variability in the amplitude spectra of signal was maintained in the population auditory midbrain neurons responses and population auditory midbrain neurons responses.

Laurel H. et al (2015) proposed a speech coding mechanism of some kind of inferior colliculus neurons for vowel formants and believed that a combination of inhibition and excitation response of midbrain neurons population encoded complex sound simultaneously. We observed that complex harmonics in temporal structure can trigger phase-locking response of multiple parts of neurons in midbrain in the context of animal model, thus led to different phase-locking intensity of neuron population. In 1968, *Worden et al.* captured a kind of evoked potential called after frequency following response (FFR) in cochlear nucleus of cat for the first time. It was much close to acoustical temporal fine structure. Some studies was executed in FFR to explore the origin of it. It is generally believed that FFR is mainly evoked in inferior colliculus (IC), which is the representation of population phase-locking encoding of auditory midbrain neurons (*Smith et al.* (1975), *Sohmer et al.* (1977), *Chandrasekaran et al.* (2009))

This paper intends to study effects of temporal modulation coding characteristics occurring in auditory midbrain on speech recognition. Based on the foundation that (1) FFR is the representation of population phase-locking encoding in auditory midbrain; (2) FFR preserves spectra-temporal features of complex sounds; (3) some complex harmonics have found to evoke phase-locking response in different parts of midbrain neurons, we propose a hypothesis that phase-locking response evoked by different group of midbrain neurons would have an effect on speech recognition. Furthermore, complex harmonics synthesized by identical starting phase would enhance the intelligibility of central neural system whereas random starting phase would weaken the intelligibility. This paper will examine the relationship between different complex harmonics and speech recognition in Mandarin Chinese. Complex harmonics is produced with identical spectral features but different temporal fine structure in a period. Those spectra-temporal characteristics are able to manipulate different group of midbrain neurons encoding speech information. Vowels in Mandarin speech are replaced by complex harmonics to produce processed speech. We play these stimuli to listeners and verify the hypothesis that we proposed.

II. METHODS

A. Subjects and Materials

Twenty young normal hearing volunteers (10 females and 10 males) who are native speakers of mandarin Chinese participated this experiment. Their ages range from 18 to 25. They are students (undergraduate or postgraduate) totally coming from Liaoning Technical University.

Sentence materials we used in this experiment are randomly chosen from Mandarin Speech Perception (MSP) corpus developed by *Fu et al.* in 2011. This corpus includes 100 sentence in total and each is designed with phonetical balance.

Each sentence in the MSP corpus is composed of seven monosyllabic words with the structure of C-V. There are 35 vowels, 21 consonants and 5 tones and pinyin rule is followed by international standard scheme of Chinese Phonetic Alphabet.

The distribution of all of these are compatible with commonly used Chinese characters. Each speech signal in the MSP corpus is segmented into three parts (consonant, vowel and silence) manually that we download from website: http://www.speech.hku.hk/MSP_VC_phn/MSP_VC_phn.html.

B. Signal Processing

The signal processing strategy in this paper divides into three parts. First of all, the pitch of every steady-like vowel segment was detected using short-time autocorrelation function method,

short-time average magnitude difference function method, and combination of above two respectively after pre-emphasis. The median of three vectors was computed so as to obtain more accurate pitch. What's more, we transfer to spectral domain, and extract magnitude and phase value of fundamental frequency and harmonics of each steady-like vowel segment. Then they were used as vowel feature parameters to resynthesize vowel signal according to equation (1). Finally, unprocessed vowel segments were replaced by corresponding synthesized vowel segments in each sentences to manufacture sentences that we would test.

$$y = \sum_{l=1}^{N} A_l \sin(2\pi * l * f_0 + \varphi) \qquad (1)$$

From (1), N represents the maximum of harmonic order (Considering the range of speech frequency and fundamental frequency of stimuli, N was chosen as 30), f_0 represents the fundamental frequency of sentence, l represents the harmonic order, and φ represents the starting phase of each sinusoidal signal.

More specifically, a method that we called three-section processing strategy was adopted to deal with vowel segment. The maximum of segment duration is 0.25s. Three-section processing strategy would process vowel center, vowel onset and vowel offset respectively so as to maintain center-onset-offset structure of vowel segment. A duration of 0.05s with energy maximum of each vowel segment was picked out as vowel center to detect vowel pitch. The former part and residual part were chosen as vowel onset and vowel offset respectively. And then rebuilding procedure was implemented in three sections. Last but not least, energy normalization for resynthesized vowel segment was made in case of inconsistent total energy effects.

Four different test conditions are designed to examine the relationship between speech intelligibility and spectra-temporal modulation encoding of auditory midbrain. Four complex harmonics with divergent temporal fine structure but identical spectra are synthesized in this experiment. They are named re_signal (vowel segment was re-synthesized by original amplitudes and phases of fundamental frequency and harmonics extracted from spectrum), sp_signal (vowel segment was resynthesized by original amplitudes extracted from spectrum of fundamental frequency and harmonics and the identical starting phase), rp_signal (vowel segment was re-synthesized by original amplitudes of fundamental frequency and harmonics extracted from spectrum and alternative starting phase of 0 and $\pi/2$), rr_signal (vowel segment was re-synthesized by original amplitudes of fundamental frequency and harmonics extracted from spectrum and random starting phase ranged $(-\pi, \pi)$ generated from randomizer) respectively . One sentence from MSP sentences materials is

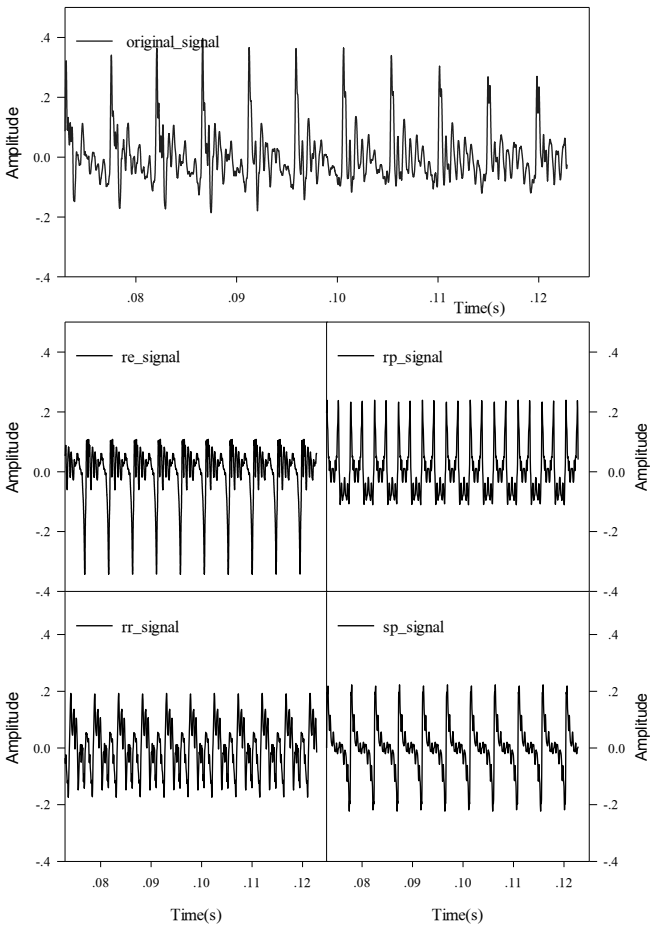

Fig.1 Five temporal waveforms of picked vowel in five conditions. re_signal denotes signal synthesized by amplitude and original starting phase of F0 and harmonics. rr_signal transfers into random starting phase. rp_signal transfers into alternative starting phase. sp_signal transfers into same starting phase.

picked out as an example, who reads ‘ [/děng huì ér ná xiē bǐng gān chī/]’ in Mandarin Chinese, which translates ‘Wait for a while to eat some cookies’ in English. Fig.1 exhibits five temporal waveform of a steady state of the vowel segment /ěng/ in picked sentence. Four processed waveforms contain essential acoustic features of original vowel and identical magnitude spectrum structure but divergent temporal fine structure. At the same time, Consideration of the origin source of FFR, four divergent temporal fine structures can trigger various frequency-following response in cerebral cortex. And thus get the goal of manipulating phase-locking response in different parts of auditory midbrain neurons for identical semantic meaning of vowel segment. Finally, those stimuli are played to human objectives to verify what the relationship between speech recognition and temporal coding characteristics of auditory midbrain.

C Procedure

The experiment was proceeded in an audiometric room located at UILab in Liaoning Technical University. Test Platform of Speech Recognition was designed to conduct the whole experiment process. Every stimuli was played to listeners at a relatively comfortable sound level and allowed to play twice at maximum. Before the formal testing, every listener was guided into an audition procedure, within which every participant could listen 10 sentences one by one in only one condition chosen by random and conducted to repeat the sentence or words that they could recognize clearly and then was given a feedback of the sentence for purpose of being familiar with test environment and process. Each participant attended an amount of 5 test conditions in formal testing and each condition consists of 10 test sentences picked randomly from MSP corpus. Within-subject method was used and no sentence was played repeatedly across all test conditions (1 audition + 5 testing). Word accuracy and sentence accuracy were recorded within experiment to quantify speech intelligibility of each condition.

III. RESULTS

Fig.2 displays the mean score of word and sentence accuracy in each test condition for mandarin speech recognition respectively. Specifically, in Fig.2, the original speech stimuli always got gorgeous score of 100 across each listener, while the average of word and sentence accuracy in other four were lower than 100. In terms of word accuracy,

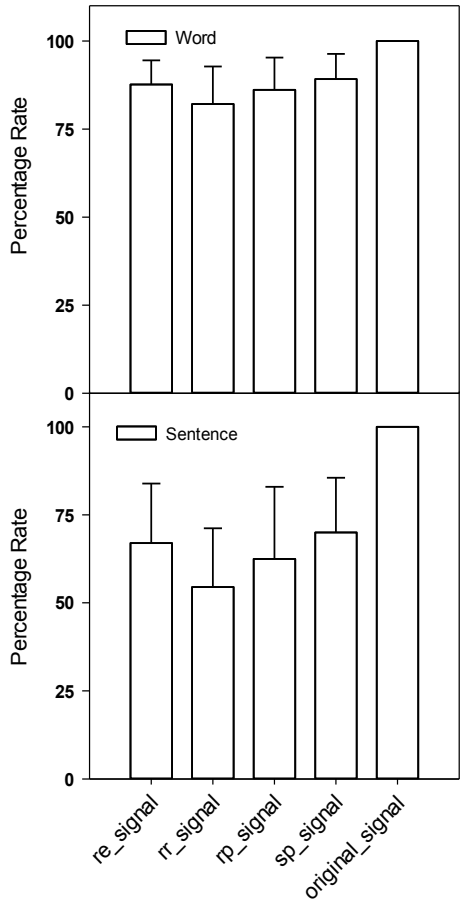

Fig.2 The mean score of word and sentence accuracy in five conditions separately. re_signal denotes signal synthesized by amplitude and original starting phase of F0 and harmonics. rr_signal transfers into random starting phase. rp_signal transfers into alternative starting phase. sp_signal transfers into same starting phase. original_signal represents unprocessed signal. The error bar denotes standard deviation of the mean.

re_signal got much better score of 87.65, while rr_signal had a worst performance of 82.10. And other two got mean score of 89.20 (sp_signal) and 86.15 (rp_signal) respectively. Look at the accuracy of sentence in four processed conditions, re_signal also got the best performance (67.00%), rr_signal was the worst

performer (54.50%) as well, sp_signal and rp_signal got mean score of 70.00% and 62.50% separately.

The paper makes different starting phase of pitch and harmonics as the within-subject factor proceeding speech recognition experiment in continuous mandarin speech. The accuracy of word and sentence are recorded as evaluation criteria for speech intelligibility. One way analysis of variance (ANOVA) with repeated measures is run in SigmaPlot software to test statistically significant difference on word and sentence accuracy across all objects and conditions. Equal variance test results in word accuracy ([F(19.76) = 18.060, P < 0.001]) and sentence accuracy ([F(19.76) = 30.966, P < 0.001]) suggesting that there is a statistically significant difference among five conditions. All pairwise multiple comparison procedures among five conditions in word and sentence accuracy with Holm-Sidak method are executed and statistical significance level is set at P<0.050 (α=0.05). Comparison results above two point that the mean difference of per treatment group (re_signal, rr_signal, rp_signal and sp_signal) versus matched group (original signal) is statistically significant, and two pairs of treatment group ([re_signal versus rr_signal] and [sp_signal and rr_signal]) have statistically significant mean difference. Specifically, in word accuracy pairwise comparison, P(original_signal vs rr_signal) < 0.001, P(original_signal vs rp_signal) < 0.001, P(original_signal vs sp_signal) < 0.001, P(original_signal vs re_signal) < 0.001, P(sp_signal vs rr_signal) = 0.012, and in sentence accuracy pairwise comparison, P(original_signal vs rr_signal) < 0.001, P(original_signal vs rp_signal) < 0.001, P(original_signal vs sp_signal) < 0.001, P(original_signal vs re_signal) < 0.001, P(re_signal vs rr_signal) = 0.004, P(sp_signal vs rr_signal) = 0.029.

IV. DISCUSIONS

Results that the mean score of all four conditions in speech recognition is lower than original signal and difference among 4 pairs is all statistically significant, suggest that feature encoded vowel segment has been weak encoded in midbrain neurons due to a shortage of some other reliable information for distinguishing vowel contrasts in speech recognition. Mandarin Chinese is a special language with lexical tone added into vowels to discriminate different meanings of a same phoneme structure, which may increase much difficulties to recognize word or sentence in this experiment. This is a subject that needs to be validated in the future.

In part II.B, we achieve the purpose of manipulating

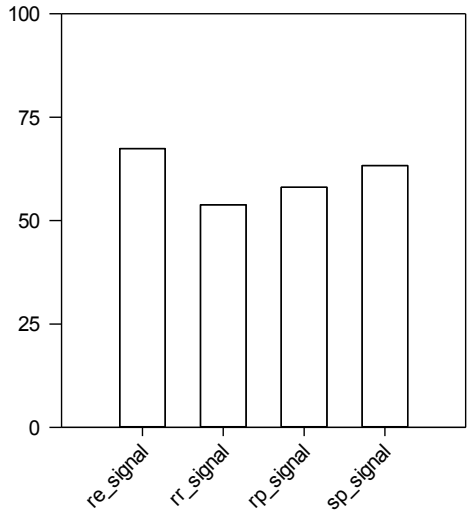

Fig.3 The mean of word_score_1 in four test conditions. re_signal denotes signal synthesized by amplitude and original starting phase of F0 and harmonics. rr_signal transfers into random starting phase. rp_signal transfers into alternative starting phase. sp_signal transfers into same starting phase.

temporal modulation coding in auditory midbrain by synthesizing different complex harmonics. Recognition results in Fig.2 imply that different temporal modulation coding characteristics in midbrain neuron population for same meaning of vowel have an effect on speech recognition in Mandarin speech whether word or sentence recognition. And same starting phase information of frequency could improve speech recognition significantly and random starting phase information of frequency could reduce word and sentence recognition on average. Auditory midbrain contains a mess of auditory neurons, part of which encoded different kinds of auditory information. The integrated response of neuron population contributes to speech comprehension. Contrast with other three complex harmonics, sp_signal has more regular temporal fine structure that may be able to bring out stronger

spiking rate contrast or phase locking response in auditory midbrain. More uniform temporal fine structure of each vowel (such as rr_signal) may would not cause strong enough neural spiking rate to trigger phase locking response of midbrain neurons in order to discriminate vowel space more easily. These results can provide some enlightenment or guideline for speech synthesis and speech enhancement. Results over experiment imply that different temporal modulation coding in midbrain neuron population may enhance or weaken some neural encoding process in advanced structure such as cerebral cortex so that influence speech intelligibility of sounds as well. So consideration of phase information is advised to add into study of encoding strategy for cochlear implant and this may become a thinkable signal processing strategy need to be verified to enhance speech intelligibility in following works.

Even though identical signal processing strategy in this experiment is executed within each vowel, the difference of listening experience and various sentence component across each listener may have an effect on speech intelligibility. In order to find out what the relationship between processed test conditions and the accuracy of largely influenced words, a further analysis of word and processed test conditions is investigated. Across all recording data, tested sentences' average sentence score (computed by times of right divided by times of testing in all conditions) named sentence_score_1 that less than 68.50% or so is picked out. Note that these sentences consist of one or two words much more uncommonly used in daily conversations, so there may be much more difficult to distinguish them from processed sentence, therefore affecting word and sentence accuracy in this experiment. Within those picked sentences, word_score_1 (similar computing method with sentence_score_1 except evaluation criteria changed as word) in each sentence across all testing samples is computed and the word whose word_score_1 less than 85% is picked out to execute word_score_1_op (computed by times of right divided by times of testing in every processed tested condition). Gathering all word_score_1_op in each condition, we collect 179 word_score_1_re_signal, 166 word_score_1_rr_signal, 148 word_score_1_rp_signal and 171 word_score_1_sp_ signal in total, and the averages of these four are displayed at Fig.3. In Fig.3, it is clear that across all test conditions, words and sentences, the average of picked word score of rr_signal (53.86) is lowest, followed by rp_signal (58.06) and sp_signal (63.33) orderly, and re_signal (67.40) is best. Such trend of average score is consistent with Fig.2 to some degree that we can believe that different temporal fine structure of same vowel have an similar effect on speech comprehension whether at total word or picked word.

V. CONCLUSIONS

In summary, this paper makes conclusions that (1) degraded temporal fine structure in speech mitigates speech intelligibility even in quiet; (2) starting phase information of frequency used to synthesize vowel have an effect on speech recognition of synthetic vowels, especially, random starting phase gets a decreased speech recognition than re_signal and identical starting phase gets a better speech intelligibility performance than re_signal and the performance of alternative starting phase information falls in between these two conditions; (3) speech temporal fine structure coding characteristics of auditory midbrain makes a difference to central neural system about discriminate some vowel signal clearly. Specifically, cophasal temporal fine structure encoding characteristics in auditory midbrain is able to capture more useful information for speech recognition in ascending central neural system and random phase temporal encoding in auditory midbrain would leave some important information facilitate speech intelligibility out.

REFERENCES

[1] 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.,2007, 122,2365-2375.

[2] Chandrasekaran B, Kraus N. "The scalp-recorded brainstem response to speech: Neural origins and plasticity". Psychophysiology, 2009.

[3] 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, 1996, pp853-856.

[4] Chen, F., Lena L. N. Wong, and Eva Y.W.Wong. "Assessing the perceptual contributions of vowels and consonants to Mandarin sentence intelligibility." in J. Acoust. Soc. Am., 2013, 134 (2), EL178-EL184.

[5] Fu, Q.J., Zhu, M., and Wang, X. S. "Development and validation of the Mandarin speech perception test." in J.Acoust. Soc.Am., 2011, 129, EL267-EL173.

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

[7] Laurel H C, Jiashu Li, Tianhao Li, and Joyce McDonough. "Using a computational model for the auditory midbrain to explore the neural representation of vowels." in The Journal of the Acoustical Society of America, 2013, volume 19.

[8] Laurel H C, Tianhao Li, Joyce McDonough. "Speech coding in the brain: representation of vowel formants by midbrain neurons tuned to sound fluctuations 1, 2, 3" in eNeuro, 2015.

[9] Smith J C, Marsh J T, Brown W S. "Far-field recorded frequency-following responses: evidence for locus of brainstem sources." Electroencephalography and Clinical Neurophysiology, 1975, 39(5):465-472.

[10] Sohmer H, Pratt H, Kinarti R. "Sources of frequency following responses (FFR) in man." Electroencephalography and Clinical Neurophysiology, 1977, 42(5): 656-664.

[11] Worden F G, Marsh J T.. "Frequency-following (microphonic-like) neural responses evoked by sound." in Electroencephalography and Clinical Neurophysiology, 1968, 25(1): 42-52.