INTERSPEECH 2014

Relative importance of AM and FM cues for speech comprehension: Effects of speaking rate and their implications for neurophysiological processing of speech

Guangting Mai [1]

[1] Department of Psychology, The University of Sheffield

gmai1@sheffield.ac.uk

Abstract

Previous studies have shown that slowly-varying amplitude modulations (AM) are crucial for speech comprehension. Moreover, recent neurophysiological studies showed low-frequency neural oscillations (< 10 Hz) are taking roles in tracking such critical AM cues which facilitate speech comprehension. However, many of such studies neglected the detailed spectral information (frequency modulations (FM)). The current paper conducted a behavioral experiment to study the relative importance of AM and FM cues for sentence intelligibility based on the hypothesis that such importance is modulated by speaking rate. By measuring the intelligibility of Mandarin sentences with selective removal of AM cues at particular AM rates or replacing FM cues with Gaussian noise, the current study found: (1) at a low speaking rate (4-Hz syllable rate), FM cues and high-rate AM cues made only marginal contributions to speech intelligibility, which is consistent with previous findings; (2) at high speaking rates (6- and 8-Hz syllable rates), however, FM cues made significant contributions even greater than AM cues at the rates suggested to be essential by previous neurophysiological studies. This result thus illustrates the relative importance of AM and FM cues at different speaking rates in Mandarin and implications for the neurophysiological speech processing were further discussed.

Index Terms: amplitude modulations (AM), frequency modulations (FM), speech intelligibility, syllable rate

1. Introduction

A large body of studies investigating speech comprehension, especially those for the purpose of cochlear implant encoding with vocoder simulations in healthy, normal-hearing subjects have shown that slowly-fluctuated temporal envelope information (amplitude modulations or AM hereafter) are playing essential role in recognizing speech signals (e.g., [1][2][3]). These studies showed that in quiet environment, preserving these AM cues in even sparsely distributed frequency channels is sufficient for speech recognition. This thus reflects the nature of redundancy of speech acoustic properties in comprehension that at least under quiet environment, neither high-rate AM components which maintain high temporal resolutions nor the spectral details (i.e., frequency modulations (FM)) within each frequency channel are essential for comprehension.

Moreover, recently growing number of neurophysiological studies have found that low-frequency neural oscillations (typically < 10 Hz) are taking the role in tracking the essential AM cues of the incoming speech signals and such role might underpin speech intelligibility which possibly serves as one of the neural mechanisms of speech processing in the brain [4][5][6]. Both [4] and [5] used magnetoencephalography (MEG) found that the low-frequency MEG phase locking to the corresponding low-rate AM components of the speech signals is correlated with the speech intelligibility. [6] further found that when artificially removing the essential AM cues (e.g., AM rates of 2 ~ 9 Hz), the previously observed neural tracking was reduced, accompanied by the reduction in speech intelligibility. Such neurophysiological studies are quite interesting since they not only gave support to the previous behavioral findings from the neural aspect, but also provided a route to find out the possible underlying mechanisms in the brain.

However, it has been recently argued that the neural-speech AM relationship may be overemphasized [7]. One of the concerns that the current paper is interested in is that many of these studies are focusing only on the AM cues but neglecting the spectral details (i.e., FM cues). From the behavioral aspect, as previously mentioned, FM cues are not necessarily needed for speech recognition in quiet environment [1][3][8]. However, FM cues are still taking an important role. For example, they can significantly improve intelligibility in noisy environment [9]. They are also important in lexical tone recognition in tonal languages like Mandarin [10]. And even in quiet environment, simply preserving the AM cues was found to be insufficient for recognition of speech at particular speaking styles, especially at high speaking rates [11]. From the neurophysiological aspect, although neural correlates of FM were previously investigated in auditory experiments [12], no such correlates were studied in speech perception [7]. For instance, in [4], only neural correlates of AM cues were considered. In [5] and [6], the FM components of the experimental stimuli were artificially replaced by Gaussian noise as many cochlear implant simulations did, which may not reflect the speech perception in the real-world scenario in normal-hearing subjects.

Based on these discussions on the role of AM and FM cues for speech comprehension, the current study aims to further clarify the roles of AM and FM cues by comparing the relative importance between them. The current study conducted a behavioral experiment using Mandarin spoken sentences as stimuli. According to [11], the intelligibility of noise-vocoded Mandarin speech was found to be worsened with increasing speaking rate, indicating that FM cues become more important at high speaking rates. It can be thus hypothesized that the relative importance of AM and FM cues is different across speaking rates. To test this hypothesis, the current study used Mandarin speech produced at 3 different speaking rates and measured the intelligibility of speech with either selective removal of AM cues at particular AM rates or using Gaussian noise in substitution for the original FM cues (see Part 2 for details).

The current study predict the results will provide insights for future neurophysiological investigations when studying speech perception in normal-hearing subjects, that if the relative importance of AM and FM cues for speech intelligibility is different across speaking rates, the neural

14 – 18 September 2014, Singapore

correlates of AM and FM cues that underpin speech intelligibility may also be modulated by speaking rates. This would therefore provide new plausible hypotheses to further study the underlying mechanisms of how the brain processes acoustic information in speech signals.

2. Materials and signal processing

2.1. Subjects and materials

Eighteen native Mandarin subjects who are normal-hearing and have no report of any hearing or speech disorders were recruited in the Capital Normal University in Beijing, Mainland China. The subjects were all undergraduate or graduate students aging from 18 to 25 years old.

During the experiment, subjects were required to listen to Mandarin sentences, which were Semantically Unpredictable Sentences (SUSs) produced by a native male speaker (F0s dynamically ranging from 80 Hz to 240 Hz) at three average syllable rates of 4 Hz, 6 Hz and 8 Hz. SUSs are sentences that are syntactically acceptable but semantically anomalous [13]. The purpose of using SUSs was to reduce the possibility that subjects guessed the content of each sentence on the basis of the semantic context, and force the subjects to focus on the words and syllables they heard. Each SUS consisted of four disyllabic (2-character) words. All the words were manually chosen from the 9,000 most frequent items of a contemporary Mandarin word frequency corpus [14]. For instance, an SUS "地图突破严肃的地点" has four disyllabic (2-character) words "地图" ("map"), "突破" ("break through"), "严肃" ("serious") and "地点" ("place"), while "的" is the particle without substantive meaning.

2.2. Signal processing

All SUSs were re-synthesized following the procedure illustrated in Fig. 1. Each sentence was band-pass filtered (BPF) into 22 frequency channels at logarithmic scale ranging from 100 to 5500 Hz. AM and FM components were then extracted from the original speech in each channel and combined together to form the reconstructed signal. The AM components were Hilbert envelopes extracted through Hilbert transform (HT) and further filtered to obtain the target AM. The FM components were extracted via the FAME algorithm developed by Zeng and colleagues [9][10]. The FM rate was then limited to 50 Hz in each frequency channel. The reason for using this algorithm rather than treating the Hilbert fine structure (HFS) as the FM is that HFS has been reported to reconstruct the envelopes (i.e., AM components) of the original speech [15]; on the other hand, the extracted FM components via FAME are acoustically independent from the AM components avoiding the envelope reconstruction [15], which therefore serves the current purpose of investigating the independent importance of AM and FM components in speech.

Since current study aims to find out how relatively important the AM and FM components are and how the importance is different across different speaking rates, the experiment stimuli were processed with 5 types of manipulations for each syllable rate. As stated in the Introduction, the way is to either artificially remove the AM components at particular AM rates or substitute Gaussian noise for the FM components and then compare the speech intelligibility in different conditions in order to achieve the purpose. The 5 manipulations are as follows and summarized in Table 1:

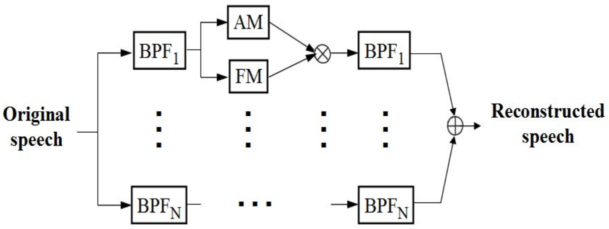

Figure 1: *Signal processing procedures based on the FAME algorithm in [9] and [10]. All signals were divided into 22 log-scale frequency channels, i.e., N = 22.*

For Manipulation 1 to 4, the FM components in each frequency channel were extracted from the original speech and then combined with the filtered AM components.

(1) Manipulation 1, which was a control condition. For AM components in each frequency channel as shown in Fig. 1, AM rates below 32 Hz, 48 Hz and 64 Hz were conserved for 4-, 6- and 8-Hz syllable rate SUSs, respectively. As stated above, the FM components were extracted from the original speech and combined with the AM components in each channel. The reconstructed speech therefore contains most crucial information (i.e., AM and FM) in the original speech which led to very high comprehensibility of over 95% syllable accuracy (see Part 4).

(2) Manipulation 2, which conserves AM rates of 4 ~ 32 Hz, 6 ~ 48 Hz, 8 ~ 64 Hz (also conserves the DC-component (0 Hz)) for 4-, 6- and 8-Hz syllable rate SUSs, respectively. In other words, the AM rates at the supra-syllabic rate (e.g., < 4 Hz for 4-Hz syllable rate SUSs) were artificially removed.

(3) Manipulation 3, which conserves AM rates of 0 ~ 4 and 8 ~ 32 Hz for 4-Hz, 0 ~ 6 and 12 ~ 48 Hz for 6-Hz, 0 ~ 8 and 16 ~ 64 Hz for 8-Hz syllable rate SUSs, respectively. In other words, the AM rates at the low sub-syllabic rate (e.g., 4 ~ 8 Hz for 4-Hz syllable rate) were artificially removed.

(4) Manipulation 4, which conserves AM rates below 8 Hz, 12 Hz and 16 Hz for 4-, 6- and 8-Hz syllable rate SUS, respectively, i.e., removing AM rates at the high sub-syllabic rate (e.g., > 8 Hz for 4-Hz syllable rate).

(5) Manipulation 5, in which the FM components were not extracted from the original speech but substituted for by Gaussian noise. The Gaussian noise was firstly filtered into respective logarithmic channels mentioned previously. It was then Hilbert transformed in each channel to obtain its Hilbert fine structure (HFS). The noise HFS was then modulated with the AM components extracted from the original speech in the same channel to form the reconstructed speech. The AM components were the same as in Manipulation 1, i.e., most crucial AM cues were conserved.

In summary, Manipulation 1 conserved most of the important acoustic information contained in the original speech, i.e., no critical AM or FM cues were absent. Manipulation 2 to 4 selectively removed particular AM cues by filtering out AM rates at either supra-syllabic or sub-syllabic rates in each frequency channel, whilst the FM cues were conserved. Manipulation 5 conserved the AM cues as in Manipulation 1 but replaced the FM cues with Gaussian noise in each channel. Therefore, as the current paper previously explained, the relative importance of AM cues at different rates and FM cues was then assessed via measuring the

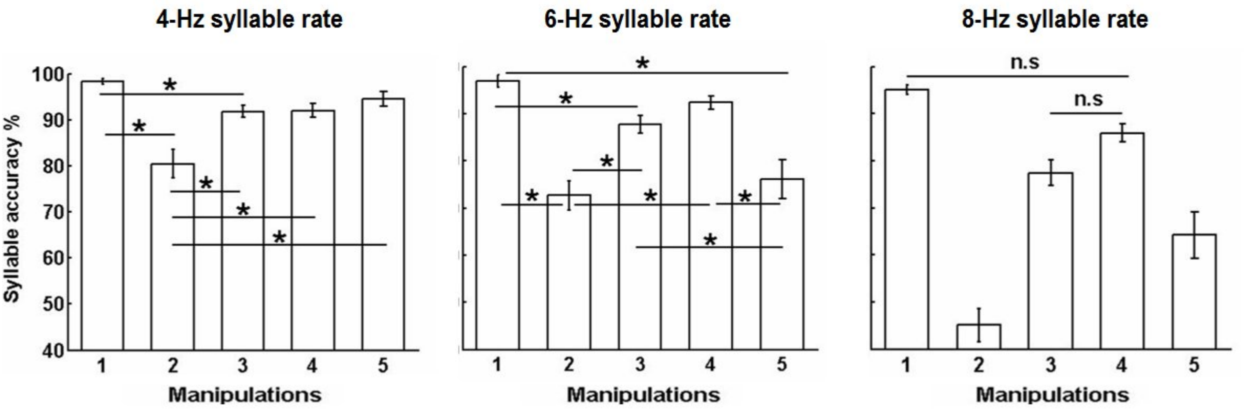

Figure 2: *Syllable identification accuracies averaged across all subjects, as the function of the 5 manipulations, grouped by the 3 different syllable rates. Error bars denote the SEM across subjects. All significances were determined by the significance level at 5% with Bonferroni correction. For the 8-Hz syllable rate, only non-significant pairs are marked.*

intelligibility of speech signals with selective exclusions of AM/FM cues by these 5 manipulations. It is further worth emphasizing that Manipulation 2 and 3 removed the conventionally critical low-rate AM components (at supra- and sub-syllabic rates respectively) suggested in previous neurophysiological studies [4][5][6], whilst the detailed temporal and spectral cues were excluded in Manipulation 4 (high-rate AM) and Manipulation 5 (FM cues), respectively.

All stimuli were processed via Matlab 7.1 (The Mathworks Inc.) and finally adjusted to the same RMS intensity.

Table 2. *Manipulations designed in the current experiment. See text for detailed descriptions*

Syllable rate	Manipulation	AM components	FM components
4 Hz	1	0 ~ 32 Hz	Extracted from speech
	2	4 ~ 32 Hz	
	3	(0 ~ 4) + (8 ~ 32) Hz	
	4	0 ~ 8 Hz	
	5	0 ~ 32 Hz	Gaussian noise
6 Hz	1	0 ~ 48 Hz	Extracted from speech
	2	6 ~ 48 Hz	
	3	(0 ~ 6) + (12 ~ 48) Hz	
	4	0 ~ 12 Hz	
	5	0 ~ 48 Hz	Gaussian noise
8 Hz	1	0 ~ 64 Hz	Extracted from speech
	2	8 ~ 64 Hz	
	3	(0 ~ 8) + (16 ~ 64) Hz	
	4	0 ~ 16 Hz	
	5	0 ~ 64 Hz	Gaussian noise

3. Tasks and experimental procedure

A set of 150 SUSs in different contents were created for the formal testing session, in which there were 15 conditions (3 syllable rate × 5 Manipulations) and 10 SUSs for each condition. All subjects were presented with this same set of SUSs. To reduce the possible biases caused by the contents of the sentences (e.g., different word frequencies), subjects were divided into 9 groups (2 subjects each) and in each condition different groups were presented with different subsets of SUSs. This was achieved through the signal processing (Part 2.2) preparing the sentence stimuli prior to the experiments. During the audio presentation, the SUSs were divided into 5 successive blocks, where each block had 30 SUSs containing SUSs of every condition (2 SUSs for each condition) and all SUSs were in an intermixed order.

Subjects were seated in a quiet room and the stimuli were presented diotically. They were instructed to write down the words or syllables they heard in each SUS on an answer sheet after it was played. To minimize the homophone confusions, they were required to transcribe what they heard using "*Pinyin*" (the official phonetic alphabet system transcribing pronunciations of Chinese characters into Latin scripts) if they could only recognize one syllable within a word. Before the formal test, each subject attended a 20- to 30-minute training session consisting of an extra set of 20 SUSs (all different from those in the formal test). Each training SUS was presented more than once with feedback of the correct answer so that subjects could get familiar with the experiment beforehand. In the formal test, each SUS was played only once with no feedback provided. The entire experiment lasted for around 2 hours.

4. Results

The speech intelligibility was calculated as syllable identification accuracies as shown in Fig. 2, which refer to the percentage of correct syllables recognized by the subjects. The accuracies were grouped by different syllable rates (4-, 6- and 8-Hz) as the function of the 5 manipulations mentioned above.

First of all, as predicted previously, Manipulation 1, which was the control condition where no essential AM or FM cues were absent, shows highest accuracy scores of over 95% for all the three syllable rate conditions. As shown in Fig.2, statistic analyses with Bonferroni correction (for multiple comparisons on p-values with the significance threshold of 5%) were further conducted for each syllable rate condition between different manipulations. The results showed that:

(1) For the 4-Hz syllable rate SUSs, consistent with previous studies, there was no significant difference between the control Manipulation (Manipulation 1) and the manipulation with removal of high-rate AM (Manipulation 4) or with replacing FM by Gaussian noise (Manipulation 5). Furthermore, as predicted, significantly higher scores were

found comparing Manipulation 1 and manipulations with removal of low-rate AM cues (Manipulation 1 vs. 2 and 1 vs. 3) and significantly higher scores for manipulations with removal of high-rate AM or FM than those with removal of low-rate AM cues (Manipulations 2 vs. 4 and 2 vs. 5).

(2) For the 6-Hz syllable rate SUSs, the main difference with the 4-Hz syllable rate SUSs is that Manipulation 5 which substituted Gaussian noise for the FM cues was found to cause significant drop in intelligibility. The accuracy for Manipulation 5 is not only significantly lower than Manipulation 1, but also significantly lower than accuracies for manipulations with removal of either low-rate or high-rate AM (i.e., Manipulation 3 and 4 respectively).

(3) For the 8-Hz syllable rate SUSs, as shown in Fig. 2, only the non-significant (n.s.) pairs (Manipulation 1 vs. 4 and 3 vs. 4) are illustrated, with other pairs all being significantly different. The pattern is similar to the 6-Hz syllable rate, but different with the 4-Hz syllable rate SUSs, also mainly on the Manipulation 5 that excluded the FM cues which consequently causes significant decrease in intelligibility.

5. Discussions

5.1. AM, FM cues, intelligibility and their neurophysiological correlates

The current study was designed to investigate the relative importance of AM cues at different AM rates and FM cues for speech intelligibility in quiet environment and it was hypothesized that such importance is different across speaking rate. Syllable accuracies of Mandarin SUSs at three different syllable rates with selective removal of AM rates or replacing FM cues with Gaussian noise were calculated. It has been found that: (1) for slow syllable rate at 4 Hz, only low-rate AM cues (< 8 Hz) were critical; (2) for high syllable rates at both 6 and 8 Hz, FM cues were also found to be critical, which is even more critical than the sub-syllable low-rate AM (Manipulation 5 vs 3 in both syllable rate cases).

Previous studies have suggested that very detailed spectral information such as FM cues in quiet environment are not essential for speech intelligibility [1][16], while the current study found that this is not true for perception of speech at high speaking rates. It thus reflects that AM cues do not necessarily keep useful linguistic information intact, whereas spectral details are also important. It may also because Mandarin speech was used in the current study and FM cues have been found to be important for lexical-tone recognition in Mandarin [9][10] (to be further discussed in Part 5.2).

From the neurophysiological aspect, as discussed in the Introduction, previous studies basically focused on the neural correlates of the low-rate AM cues (< 10 Hz) that have been conventionally suggested to be essential for speech comprehension and argued that the tracking of such AM cues via brain oscillations may be the underlying mechanism in perceiving speech signals [4][5][6]. The implications of current results for the neurophysiological processing of speech will thus emphasize that the neural correlates of AM and FM cues that underpin speech intelligibility should be different across syllable rates.

There is actually a recent study by Obleser *et al.* [17] using fMRI investigating the temporal and spectral information of speech and their impacts on intelligibility that co-varied with the brain responses in bilateral superior temporal sulcus (STS), an area thought to be crucial for speech comprehensibility. On the temporal domain, this study selectively removed different amount of AM cues similar to the current study, whilst on the spectral domain, number of frequency channels were changed to control the spectral resolution. This study found that the intelligibility controlled by temporal manipulations of AM rates co-varied with brain responses in the left STS, while the intelligibility controlled by the spectral manipulations on the number of frequency channels co-varied with brain responses in the right STS. This study was a pioneer study investigating neural correlates of speech acoustic cues in both temporal and spectral domain. However, the current study suggests that not only the spectral resolution but also FM cues should be further considered as important information in the spectral domain, especially at high speaking rates. One further clue for this argument is that in the current experimental stimuli, all speech signals were divided into 22 logarithmic-scale frequency channels from 100 to 5500 Hz (see Part 2.2), the number of which (i.e., the spectral resolution) is high enough for speech comprehension, and yet FM cues were still found to take a critical role.

Furthermore, based on the results obtained by Obleser *et al.* [17] that temporal information (AM cues) and spectral information which underpin speech intelligibility are respectively processed in the left and right hemisphere, it is plausible to hypothesize that FM cues which are spectral details and the temporal AM cues would be processed under separate neurophysiological mechanisms.

5.2. Importance of FM cues for tonal languages and its potential neurophysiological correlates

As previously mentioned, the increasing importance of FM cues with increasing speaking rate may be related the importance of FM cues for the lexical-tone recognition in Mandarin [9][10]. It is thus worth exploring in the future if similar effect can be found with speech stimuli of a non-tonal language.

From the neurophysiological aspect, lexical tones have been found to be processed more in the right hemisphere than other linguistic features [18]. This can be compared to the results in Obleser *et al.* [17] that right hemisphere is more sensitive to spectral cues than to temporal AM cues in speech. In other words, it is worth pondering if the processing FM cues could offer any clues to the underlying neural mechanisms in the right hemisphere for lexical-tone recognition in tonal languages.

In summary, the current study thus provided evidence for the hypothesis stated at the beginning of the paper that the relative importance of the two main acoustic features in speech, AM and FM cues, is different across speaking rates. The current result argued for its neurophysiological implications based on the fact that most current neuroimaging studies about speech acoustic features that control comprehension are focusing on the AM cues only. Other issues such as the importance of FM cues in tonal languages and its potential neural correlates based on previous studies were also discussed.

6. References

[1] Shannon, R.V., Zeng, F-G., Kamath, V., Wygonski, J. and Ekelid, M., "Speech recognition with primarily temporal cues", *Science*, 270(5234): 303-304, 1995.

[2] Arai, T., Pavel, M., Hermansky, H. and Avendano, C., “Syllable intelligibility for temporally filtered LPC cepstral trajectories”, *J. Acoust. Soc. Am.*, 105: 2783-2791, 1999.

[3] Xu, L., Thompson, C. S., and Pfingst, B. E., “Relative contributions of spectral and temporal cues for phoneme recognition,” *J. Acoust. Soc. Am.*, 117: 3255-3267, 2005.

[4] Ahissar, E., Nagarajan, S., Ahissar, M., Protopapas, A., Mahncke, H., and Merzenich, M. M., “Speech comprehension is correlated with temporal response patterns recorded from auditory cortex”, *Proc. Natl. Acad. Sci.*, 98(23), 13367-13372, 2001

[5] Peelle, J. E., Gross, J., and Davis M. H., “Phase-locked responses to speech in human auditory cortex are enhanced during comprehension”, *Cerebral Cortex*, doi:10.1093/cercor/bhs118, 2012.

[6] Doelling, K. B., Arnal, L. H., Ghitza, O., and Poeppel D., “ Acoustic landmarks drive delta-theta oscillations to enable speech comprehension by facilitating perceptual parsing”, *NeuroImage*, 85:761-768, 2014.

[7] Obleser, J., Hermann, B., and Henry, M. J., “Neural oscillations in speech: don’t be enslaved by the envelope” *Frontiers in Human Neuroscience*, 6(250): 1-4, 2012.

[8] Dorman. M., Loizou, P., Rainey, D., “Speech intelligibility as a function of the number of channels of stimulation for signal processors using sine-wave and noise-band outputs”, *J. Acoust. Soc. Am.*, 102: 2403-2411, 1997.

[9] Zeng F-G., Nie, K., Stickney, G. S., Kong Y-Y., Vongphoe, M., Bhargave, A., Wei, C., Cao, K., “Speech recognition with amplitude and frequency modulations”, *Proc. Natl. Acad. Sci.*, 102(7): 2293-2298, 2005.

[10] Nie, K. B., Stickney, G.S., and Zeng F-G., “Encoding frequency modulation to improve cochlear implant performance in noise”, *IEEE Trans. On Biomedical Engineering*, 52: 64-73, 2005.

[11] Li Y., Zhang, G., Kang, H., Liu, S., Han, D., and Fu, Q-J., “Effects of speaking style on speech intelligibility for Mandarin-speaking cochlear implant users”, *J. Acoust. Soc. Am.*, 129(6), 2011.

[12] Picton, T. W., John, M. S., Dimitrijevic, A., and Purcell, D., “Human auditory steady-state response”, *Int. J. Audiol.* 42: 177-219, 2003.

[13] Benoit, C., Grice, M., and Hazan, V., “The SUS test: A method for the assessment of text-to-speech synthesis intelligibility using Semantically Unpredictable Sentences,” Speech Communication, 18: 381-392, 1996.

[14] Cheng, C. C., “Word List with Accumulated Word Frequency in Sinica Corpus,” Institute of Linguistics, Academia Sinica, Taiwan, 2005.

[15] Sheft, S., Ardoint, M., and Lorenzi, C., “Speech identification based on temporal fine structure cues”, *J. Acoust. Soc. Am.*, 124(1): 562-575, 2008.

[16] Saberi, K., and Perrott, D. R., “Coginitve restoration of reversed speech”, *Nature*, 398: 760, 1999.

[17] Obleser, J., Eisner, F., and Kotz, S. A., “Bilateral speech comprehension reflects differential sensitivity to spectral and temporal features”, *J. Neurosci.*, 28(32): 8116-8124, 2008.

[18] Klein, D., Zatorre, R. J., Milner, B., Zhao, V., “A cross-linguistic PET study of tone perception in Mandarin Chinese and English speakers”, *NeuroImage*, 13: 646-653, 2001.