ISCA Archive
http://www.isca-speech.org/archive

INTERSPEECH 2012
ISCA's 13th Annual Conference
Portland, OR, USA
September 9-13, 2012

Relative Importance of Temporal Envelope and Fine Structure Cues in Low- and High-Order Harmonic Regions for Mandarin Lexical-Tone Recognition

Guangting Mai [1]

[1] Language Engineering Laboratory, Department of Electronic Engineering,
The Chinese University of Hong Kong, Hong Kong
gtmai@ee.cuhk.edu.hk

Abstract

Importance of speech temporal envelope (TE) and fine structure (TFS) cues for lexical-tone recognition has been investigated in normal-hearing subjects [1][2]. The present study explores the relative importance of TE and TFS cues in low- (LH) versus high-order harmonic (HH) regions, using “acoustic chimeras” [3] with Mandarin monosyllables divided into 8 and 16 frequency channels that the current multichannel cochlear prosthesis can provide. The results show: (1) TE in both LH and HH regions make contributions to lexical-tone recognition without the existence of TFS of the original speech, but their relative importance is modulated by the number of channels; (2) TFS in LH region takes the major role in recognition, but is not enough for perfect performances; (3) TE in both LH and HH regions and TFS in HH region make significant but different complementary contributions based on the presence of TFS in LH region. Current results further address potential implications for cochlear implant stimulations for lexical-tones with combination of newly-developed encoding strategies [4][5][6].

Index Terms: Mandarin lexical-tone recognition, temporal envelope and fine structure, low- and high-order harmonics, acoustic chimera, cochlear implants

1. Introduction

Signals in the time domain can be mathematically decomposed into slow-varying temporal envelope (TE) and fast-varying temporal fine structure (TFS). It has been reported that TE and TFS cues are making different contributions to human speech perception [7][8][3]. [7] and [8] showed that when preserving TE cues through half-way rectification and low-pass filtering and replacing TFS cues with noise in a few frequency channels, the intelligibility of English speech is still perfectly conserved. This result was subsequently confirmed in the work by Smith *et al* [3], which decomposed signals into TE and TFS by “Hilbert Transform” (HT) and constructed a “speech-speech chimera” whose envelope belonged to TE of one signal and fine structure belonged to TFS of another within each frequency channel. It was found that sentence identification was based on TE rather than TFS cues with increasing number of channels.

Such results are important information for cochlear prosthesis and researchers have begun paying attention to such importance for lexical-tone perception in tonal languages. Studies have illustrated that without the existence of TFS of the original speech signals, TE cues can make crucial contributions to lexical-tone recognition [1]. On the other hand, using “acoustic chimeras” similar to [3], Xu and Pfingst [2] examined the relative importance of TE and TFS using Mandarin monosyllables and found that for normal-hearing Mandarin speakers, lexical-tones are recognized based on TFS rather than TE cues, which suggested that delivering TFS cues in cochlear implants (CI) may be beneficial for CI users.

However, relative importance of TE and TFS cues in different frequency regions has not been thoroughly considered, which would have potential influence in CI designs, e.g., for patients with hearing loss at particular frequencies or patients who are exposed to background noise at different frequencies. Yuen *et al*. [9] previously showed that TE and periodicity cues in the region of >1000 Hz are more important than those in the region of <1000 Hz for Cantonese lexical-tone recognition. It has also been revealed that in harmonic complex tones, pitch perception is based on TE and TFS cues differently between low-order resolved harmonics and high-order unresolved harmonics [10][11]. However, such relative importance is still not clearly investigated. The present study aims to explore the relative importance of TE and TFS cues in low-order harmonic (LH) and high-order harmonic (HH) regions for Mandarin lexical-tone recognition.

2. Experiment Method

2.1. Overview

Three experiments were conducted using re-synthesized speech materials similar to “acoustic chimera” in [3] based on naturally produced Mandarin monosyllables with the 4 different tones (level, rising, falling-rising and falling) by a native male Mandarin speaker whose average F_0 is around 140 Hz. The boundary between “LH” and “HH” region was determined at the frequency slightly above the value of the 5th harmonics of the average F_0 (724 Hz, see Table 1) based on the previous finding arguing that the frequency region below the 6th harmonics is dominant for pitch perception [12]. Furthermore, 5th or 6th harmonics are also considered as a possible transition point from being resolved to being unresolved [10]. Native Mandarin speakers took part in the 2-hour lexical-tone identification experiments. Exp. 1 examined the importance of TE cues in LH *versus* HH region without TFS of the original speech. Exp. 2 examined the importance of TFS cues in LH *versus* HH region. In Exp. 3, relative complementary contributions of TE in LH and HH regions and TFS in HH region were tested in the presence of TFS in LH region. In addition, 8 and 16 channel decompositions were used as in [2]. This choice matches current multichannel cochlear prosthesis that can provide 8 and 16 channels distributed from 80 to 8820 Hz according to the Greenwood Function, see Table 1. LH region includes the first 3 and first 6 channels for the 8- and 16-channel chimera, respectively, while HH region includes the rest of the channels.

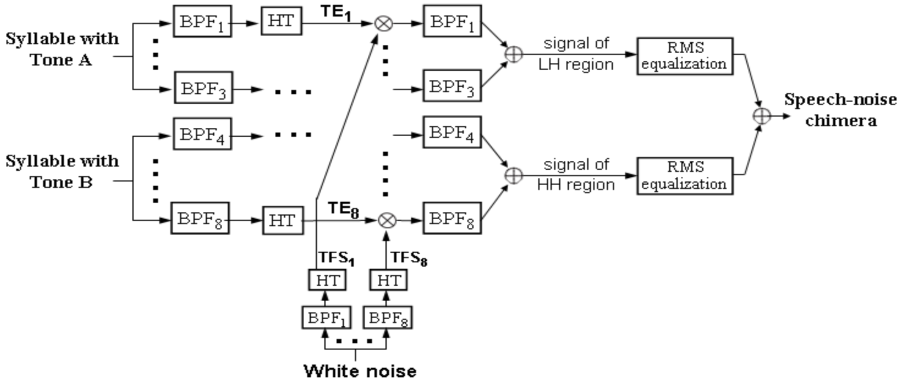

Figure 1: *Processing of the 8-channel speech-noise chimera in Exp. 1, with TFS of noise in each channel, TEs in LH region from a syllable with one tone and TEs in HH region from the same base syllable with another tone. BPF: Band-Pass Filtering; HT: Hilbert Transform*

Table 1. *Channel distributions which are the same as in [2]*

Frequency region	8 channels		16 channels	
	No.	Range (Hz)	No.	Range (Hz)
Low-order harmonic (LH) region	1	80 ~ 205	1	80 ~ 135
			2	135 ~ 205
	2	205 ~ 405	3	205 ~ 293
			4	293 ~ 405
	3	405 ~ 724	5	405 ~ 546
			6	546 ~ 724
High-order harmonic (HH) region	4	724 ~ 1236	7	724 ~ 950
			8	950 ~ 1236
	5	1236 ~ 2055	9	1236 ~ 1598
			10	1598 ~ 2055
	6	2055 ~ 3365	11	2055 ~ 2634
			12	2634 ~ 3365
	7	3365 ~ 5463	13	3365 ~ 4292
			14	4292 ~ 5463
	8	5463 ~ 8820	15	5463 ~ 6945
			16	6945 ~ 8820

2.2. Original and re-synthesized speech

The base syllables of the original speech are /ji/, /fu/, /tç^{h}ien/, /tşə/, /tuo/, /ma/ in Phonetic Alphabet, where the nucleuses cover most of the vowel space in the Mandarin vowel system. The average F_0 of the materials is about 140 Hz and the time duration is around 350 ms in the vowel portion. All the syllables correspond to real Chinese characters. Before creating the "acoustic chimeras", all materials were time-aligned at the consonant-vowel boundaries and kept the durations identical for monosyllables with the same base syllable using the time-compression/expansion algorithm PSOLA in software PRAAT.

Following abbreviations are defined: TE-LH, TFS-LH, TE-HH and TFS-HH refer to TEs extracted from channels in LH region, TFSs from channels in LH region, TEs from channels in HH region and TFSs from channels in HH region, respectively.

Speech processing procedure of the re-synthesized speech in Exp. 1 is shown as *Fig. 1*. TE-LH and TE-HH were extracted through Hilbert Transform (HT) [3] respectively from a monosyllable with one tone and that of the same base syllable with another tone. The extracted TE in each channel was then modulated by the TFS which was also extracted through HT from the broadband Gaussian noise in the respective channel. As shown in Fig. 1, the resultant chimeric signals of LH region and HH region were adjusted to retain the same RMS energy before combining them as the final speech-noise chimera. The aim of the RMS equalization is to avoid the possibility that the contribution difference between LH and HH region may be caused by their energy level difference. Further re-examination found that such RMS equalization on LH and HH regions of any originally pronounced syllable does not obviously affect the naturalness of the syllable, hence that little perceptual artifacts or biases were introduced.

The speech processing in Exp. 2 is similar to Exp. 1. There were two types of stimuli: speech-speech chimeras with (1) TFS-LH from the syllable with one tone and other TE and TFS cues (TE-LH, TE-HH and TFS-HH) from the same base syllable with another tone; (2) TFS-HH from the syllable with one tone and other cues (TE-LH, TE-HH TFS-LH) from the syllable with another tone. Such chimeras were created so as to test to what degree lexical-tones can be retrieved when only TFS cues in LH or HH region are available.

Based on Exp. 2 (see the results in *Part 3.2*), Exp. 3 were conducted with three types of stimuli: chimeras with (1) TFS-LH and TFS-HH from the syllable with one tone and other cues (TE-LH and TE-HH) from the same base syllable with another tone (the same type of stimuli in [2]); (2) TFS-LH and TE-LH from the syllable with one tone and other cues (TFS-HH and TE-HH) from another tone; (3) TFS-LH and TE-HH from the syllable with one tone and other cues (TFS-HH and TE-LH) from the syllable with another tone.

All the re-synthesized syllables were finally adjusted to the same energy level. All processing was performed in Matlab.

2.3. Subjects and tasks

23 normal-hearing Mandarin native speakers recruited in Mainland China (21 ~ 24 years of age) were instructed to listen to the re-synthesized monosyllables and give a single forced-choice selection on which tone of the syllable they heard in each trial (paper work with an answer sheet). Each type of re-synthesized syllables described in *Part 2.2* has 72 stimuli (6 syllables × 12 pairs) and each stimulus was played only once. Trials with different types of stimuli within each experiment were intermixed.

3. Experiment results

3.1. Experiment 1

Numerical data for all the three experiments are in Table 2.

Fig. 2 shows the results of Exp. 1, comparing the importance between TE-LH and TE-HH when TFS in each channel is replaced by noise. In the 8-channel case, tone responses are significantly more consistent with TE-HH (the white bar) than TE-LH (the hatched bar which is around chance level) ($p < 0.01$, Bonferroni correction). This is consistent with previous study illustrating that TEs in the high frequency region are more important than those the in low frequency region for lexical-tone recognition [9]. However, in the 16-channel case, the contribution of TE-LH increases while that of TE-HH decreases compared to the 8-channel case (no significance between the two, $p > 0.05$). It thus shows the relative importance between TE cues in LH and HH region is modulated by the number of channels used.

3.2. Experiment 2

Previous study showed that TFS cues are taking the dominant role in lexical-tone recognition rather than TE cues [2]. Exp. 2 further examined the relative importance of TFS cues in LH *versus* HH region by testing how lexical-tone can be perceptually retrieved when only TFS-LH or TFS-HH is available (see speech processing in *Part 2.2*). As illustrated in Fig. 3, lexical-tone responses is up to 87% and 65% for 8- and 16-channel cases based on TFS-LH. On the contrast, less than 1% responses were obtained based on TFS-HH (see detailed numerical data in Table 2).

This result confirms that TFS cues in LH region are taking the major role in lexical-tone recognition, which is consistent with the suggestion that pitch perception in the low frequency region is dominated by TFS cues [11]. A more useful finding is that the degree of such dominance is found to be modulated by the number of channels. The response based on the TFS-LH of the 8-channel stimuli is significantly higher than that of the 16-channel stimuli ($p<10^{-10}$).

3.3. Experiment 3

Although TFS-LH is taking the major role in the lexical-tone recognition, satisfactory response has not been achieved, especially in the 16-channel case (65%). Exp. 3 tested the respective complementary contributions of TE-LH, TE-HH and TFS-HH, based upon the presence of TFS-LH (see speech processing in *Part 2.2*). As shown in Fig. 4, obvious complementary effects can be seen for TFS-HH, TE-LH and TE-HH in both 8- and 16-channel cases (comparing responses in Exp. 3 with those solely based on TFS-LH in Exp. 2).

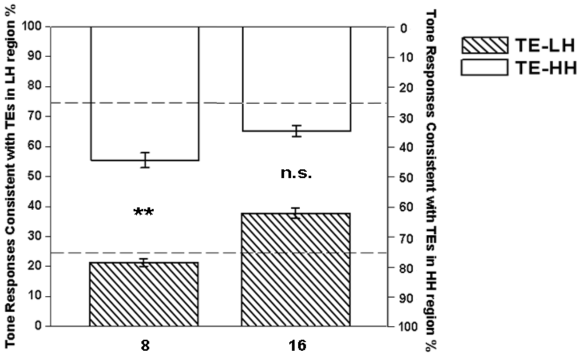

Figure 2: *Results of Exp. 1. Hatched bars represent tone responses consistent with TE-LH using the left ordinate, while white bars represent the responses consistent with TE-HH using the right ordinate. Error bars represent SEM across the subjects. Horizontal dashed lines stand for the chance levels (25%). **p<0.01; n.s.: not significant*

It is shown in Fig. 4 that: (1) TE-LH (hatched bars) makes greater complementary contributions than TFS-HH (statistically significant, $p<0.01$) and TE-HH in the 16-channel but not in the 8-channel case; (2) the contribution of (TFS-LH + TE-HH) (black bars) decreases as the number of channels increases from 8 to 16 ($p < 0.05$); same trend takes place for (TFS-LH + TFS-HH) (white bars) ($p < 0.01$), which is consistent with the results in [2]. These results thus highlight the importance of the intactness of temporal cues in LH region (TFS-LH + TE-LH) for lexical-tone recognition in the case with higher number of frequency channels, where more detailed spectral information of the original speech was provided than in the 8-channel case.

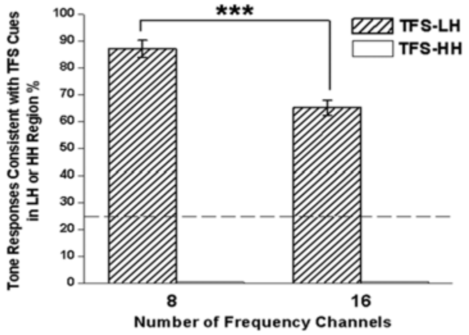

Figure 3: *Results of Exp. 2. Hatched bars represent tone responses consistent with TFS-LH while white bars represent responses consistent with TFS-HH (scores too low to be seen). ***p<10*$^{-10}$

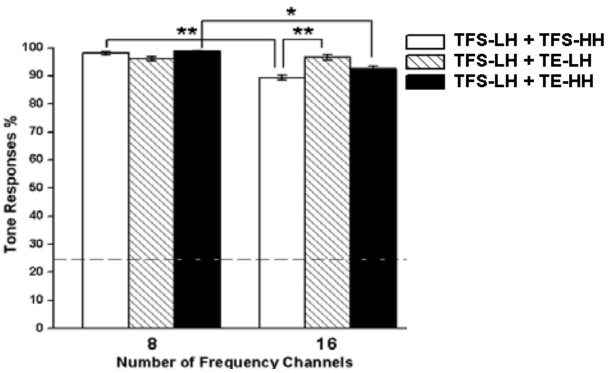

Figure 4: *Results of Exp. 3. White, hatched and black bars represent tone responses consistent with TFS-LH+TFS-HH, TFS-LH+TE-LH and TFS-LH+TE-HH, respectively.*p<0.05; **p<0.01*

4. Discussions

The current study used different types of “acoustic chimeras” to investigate the relative importance of TE/TFS cues in LH and HH regions for lexical-tone recognition. “Acoustic chimera” was employed so that subjects can base solely on particular acoustic cues with other cues being substituted.

Summarizing the results from the three experiments, we find: (1) Both TE-LH and TE-HH are making contributions to lexical-tone recognition without the existence of TFS of the original speech, and the relative importance between them is modulated by the number of channels. The importance of TE-LH increases as the number of channels increases and the other way round for TE-HH; (2) TFS-LH is taking the major role in lexical-tone recognition, but satisfactory response has not been achieved based solely on TFS-LH, especially when the number of channels increases to 16; (3) TFS-HH, TE-LH and TE-HH are making obvious complementary contributions based upon the existence of TFS-LH, and their relative contributions are modulated by the number of channels.

This study thus expands the results in [2] revealing the dominance of TFS cues in lexical-tone recognition. Here it has been further unfolded that such dominance is significantly different in different frequency regions and is modulated by the number of channels. Furthermore, complementary contributions of TE and TFS cues in different regions were investigated.

4.1. Implications for psychoacoustics

From the psychoacoustic aspect, it is interesting to consider if the differing importance between LH and HH region could reflect different mechanisms of resolved and unresolved harmonics in lexical-tone perception. The boundary between LH and HH region in the current study is the frequency slightly higher than the 5th harmonics of the average F_0 value of the original speech materials, based on the finding that the frequency region below the 6th harmonics is dominant for pitch perception of complex tones [12]. Such boundary setting is also consistent with the argument that harmonics can be resolved up to 5 to 8th harmonics with F_0 of around 100 Hz [10], indicating that harmonics in LH region in the current study are basically resolved while most harmonics in HH region are unresolved. However, since the resolvability and pitch detections were based on experiments with non-speech stimuli, harmonic resolvability and its association with lexical-tone recognition may not be necessarily transparent in speech signals due to the complex spectral structures and temporal dynamics.

In comparison with previous work, some of the current results (the 8-channel case in Exp. 1 and 2 which show the importance of TE-HH and TFS-LH, respectively) are consistent with studies that pitch in complex tones can be perceptually estimated based on TFS of low-order resolved or TE of high-order unresolved harmonics [10][11]. However, the current study also shows that all the four parameters (TE-LH, TE-HH, TFS-LH and TFS-HH) can make significant contributions, which indicates a more complicated mode of lexical-tone perception than pitch perception of non-speech stimuli.

4.2. Implications for cochlear implants (CI)

Although such acoustic simulations in normal-hearing subjects may not directly reflect the perception in CI patients, they could allow proper assessments of the roles of crucial acoustic cues. Most of the current CIs are focusing on delivering TE cues, however, some newly-developed strategies such as Fine Structure Programming [4], Frequency-Amplitude-Modulation-Encoding [5] and Harmonic Single Sideband Encoder [6], which are aiming to convey TFS cues effectively, will help us address possible applications for CIs in lexical-tone recognition.

One implication of the present study is that current results are potential for better designs and performance assessments of CI users who are exposed to background noise in different frequencies (e.g., low/high frequency noise corresponding to LH/HH region) in their daily communications, or patients with selective hearing loss at particular frequencies and with different abilities of using TE/TFS cues. Particularly, lexical-tone recognition performances can be improved through designing strategies to preserve and enhance the most essential acoustic cues. It is also important to take the observations seriously that relative contributions of different cues are modulated by the number of frequency channels.

Future work will look into the effects of different base syllables and conduct experiments with more realistic acoustic simulations under complex background environments.

Table 2. *Numerical results of the experiments; each response data was obtained through 72 different stimuli by all the subjects.*

Exp.	Acoustic cues	Channel number	Percent responses consistent with the acoustic cues%(SEM)	Fig.
Exp.1	TE-LH	8	21.3 (1.2)	Fig.2
		16	37.8 (1.7)	
	TE-HH	8	44.3 (2.5)	
		16	34.4 (1.8)	
Exp.2	TFS-LH	8	87.1 (3.2)	Fig.3
		16	65.1 (2.8)	
	TFS-HH	8	0.3 (0.2)	
		16	0.4 (0.2)	
Exp.3	TFS-LH + TFS-HH	8	98.2 (0.6)	Fig.4
		16	89.4 (0.9)	
	TFS-LH + TE-LH	8	96.2 (0.8)	
		16	96.7 (1.0)	
	TFS-LH + TE-HH	8	99.0 (0.2)	
		16	92.6 (0.9)	

5. References

[1] Fu, Q.J., Zeng, F.G., Shannon, R.V. and Soli, S.D., "Importance of tonal envelope cues in Chinese speech recognition", J. Acoust. Soc. Am., 104(1): 505-510, 1998.

[2] Xu, L., and Pfingst, B.E., "Relative importance of temporal envelope and fine structure in lexical-tone perception", J. Acoust. Soc. Am., 114(6): 3024-3027, 2003.

[3] Smith, Z.M., Delgutte, B. and Oxenham, A.J., "Chimaeric sounds reveal dichotomies in auditory perception", Nature, 416(7): 87-90, 2002.

[4] Riss, D., Amoldner, C., Reiss, S., Baumgartner, W.D. and Hamzavi, J.S., "1-year results using the Opus speech processor with the fine structure speech coding strategy", Acta Otolaryngol. 129(9): 988-991, 2009.

[5] Nie, K., Stickney G., and Zeng, F.G., "Encoding Frequency Modulation to Improve Cochlear Implant Performance in Noise", IEEE Trans. Biomed. Engineering, 52(1): 64-73, 2005.

[6] Li, X., Nie, K., Atlas, L. and Rubinstein, J., "Harmonic coherent demodulation for improving sound coding in cochlear implants", IEEE ICASSP, 5462-5465, 2010.

[7] 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.

[8] Dorman, M., Loizou P. and 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., 24: 175-184, 1997.

[9] Yuen, K.C.P., Yuan, M., Lee, T., Soli, S.D., Tong, M.C.F. and van Hasselt, C.A., "Frequency-specific temporal envelope and periodicity components for lexical tone identification in Cantonese", Ear Hear, 28(2): 107-113, 2007.

[10] Plack, C.J., Oxenham, A.J., Fay, R.R. and Popper, A.N., "Pitch: Neural coding and perception", Springer, New York, 2005.

[11] Oxenham, A.J., Micheyl, C., Keebler, M.V., "Can temporal fine structure represent the fundamental frequency of unresolved harmonics?", J. Acoust. Soc. Am., 125(4): 2189-2199, 2009.

[12] Ritsma, R.J., "Frequencies dominant in the perception of the pitch of complex sounds", J. Acoust. Soc. Am., 42(1): 191-198, 1967.