A NEW FREQUENCY LOWERING TECHNIQUE FOR MANDARIN-SPEAKING HEARING AID USERS

Yen-Teh Liu[1], *Ronald Y. Chang*[1], *Yu Tsao*[1*], *and Yi-ping Chang*[2]

[1]Research Center for Information Technology Innovation, Academia Sinica, Taiwan
[2]Speech and Hearing Science Research Institute, Children's Hearing Foundation, Taiwan
[*]Email: yu.tsao@citi.sinica.edu.tw

ABSTRACT

Frequency lowering technologies have been developed to improve the hearing experience of people with high-frequency hearing loss by preserving high-frequency information of the speech in a lower-frequency band. The technique was shown effective on the English speech for English-speaking people with severe to profound high-frequency hearing loss. The unique phonology of the Mandarin makes it worthwhile to examine the effect of this technique on the Mandarin speech, which has not been well investigated. In this paper, we propose a new frequency transposition framework that incorporates the language-specific characteristics of the Mandarin for efficient signal processing, a deep neural network (DNN) technique for Mandarin consonant/vowel classification, and the equal loudness model in psychoacoustics for time-domain compensation. Our results show that the proposed algorithm achieves statistically significant improvements in Mandarin speech recognition performance in a simulated severe high-frequency hearing loss condition.

Index Terms— Frequency lowering technique, Mandarin speech recognition, hearing aids, deep neural network.

1. INTRODUCTION

High-frequency hearing loss is a common age-related hearing loss. The inability to access the high-frequency information compromises speech perception, especially consonant perception. The traditional amplifying hearing aids can benefit people with mild-to-moderate high frequency hearing loss, but provide limited advantages for people with severe-to-profound high-frequency hearing loss [1, 2].

An alternative approach to dealing with high-frequency hearing loss is through frequency lowering technologies [3], which include various signal processing techniques such as channel vocoder, frequency compression, and frequency transposition (FT), all aiming at presenting high-frequency information in a lower frequency region that is accessible for hearing-impaired listeners. Channel vocoder divides the speech signal into frequency bands by bandpass filters and extracts the envelopes of high-frequency signals to modulate a noise source, which will be added to the unmodified low-frequency signals [4, 5]. Frequency compression reduces the bandwidth of a speech signal in a linear or nonlinear fashion [6]. FT shifts the high-frequency components to a lower-frequency region and adds them to the unprocessed lower-frequency signals [7–9]. The FT method became the first frequency-lowering technique implemented in a commercial hearing aid [3].

This paper proposes a new FT scheme for Mandarin-speaking hearing aid users, which inherits the core idea of the original FT technique yet incorporates many new insights. First, Mandarin consonants only appear in the initial position of a syllable and not in the final (or coda) position of a syllable. Hearing-impaired people therefore are likely to perceive sounds from the formant transition between high-frequency consonants and the following vowels. In fact, we have observed that the vowels are still recognizable even without the high-frequency information, and unnecessary frequency transposition may degrade vowel recognition. Thus, we adopt a deep neural network (DNN) consonant/vowel classifier and perform FT only on consonants. Second, a fixed source band in FT may not adequately represent different speech signals with varying frequency components. Thus, we propose an adaptive source band to capture the most important information in the spectrum. We also propose a weighted superposition of different frequency components based on the equal loudness model in psychoacoustics.

This paper is organized as follows. Section 2 describes the proposed method and the evaluation method. Section 3 presents the testing results and discussion. Section 4 concludes the paper and outlines future work.

2. METHODS

2.1. The Proposed Algorithm

The speech signal was sampled at 16 kHz and quantized in 16 bits. The speech signal was divided into frames of 8 ms (128 sample points) and convolved with a Hanning window to simulate the on-line processing for hearing aids. The flow chart of the proposed method is shown in Fig. 1. First, speech features are extracted from the windowed speech and fed into a Deep Neural Network (DNN) classifier for Mandarin con-

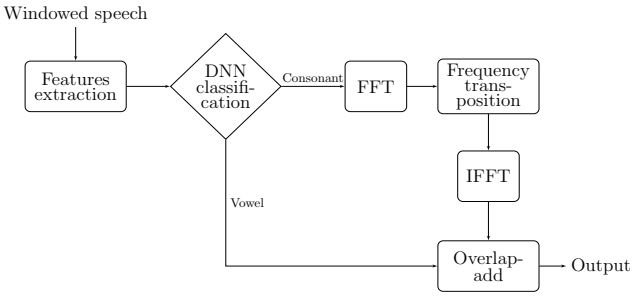

Fig. 1. Flow chart of the proposed method.

sonant/vowel classification. We implement the DNN model with 3 layers (10 nodes each layers) from [10], which reported a 93.80% accuracy in Mandarin consonant/vowel classification. If a frame is classified as a consonant, the frame is transformed to the frequency domain by the Fast Fourier Transform (FFT) and processed by the proposed frequency transposition (FT) method detailed below. If a frame is classified as a vowel, it is not processed by FT. The FT-processed consonant frames are transformed back to the time domain by the Inverse FFT (IFFT) and overlap-added with the vowel frames (64 points overlap between frames) to produce the testing speech signal.

The proposed FT scheme inherits the core idea of the original technique with a new adaptive source band and weighted superposition of frequency components. The proposed method is described as follows.

Target band: The target band in FT should be carefully determined. If the target band is too low, transposition might disturb the existing low-frequency components and negatively affect hearing for users with residual hearing at low frequencies. If the target band is too high, transposition is not meaningful since it will not preserve high-frequency information at sufficiently low frequencies that hearing-impaired listeners might be able to access. We set the target band to be 750–1500 Hz. 1500 Hz is the cut-off frequency in our experiments to simulate a severe degree of hearing loss. 750 Hz is chosen so that much of the vowel information and the natural pitch of the speaker are preserved since frequencies below 750 Hz are left unchanged.

Source band: A fixed source band in FT [9, 11] may not adequately represent different speech signals with varying frequency components. Thus, we propose an adaptive source band on a frame-by-frame basis. It was shown [12, 13] that the first spectral moment M_1, which can represent the centroid of the spectrum, well characterizes the consonants. M_1 is calculated as

$$M_1 = \frac{\sum_{l=1}^{\frac{N}{2}} P(l)(f_s \times l/N)}{\sum_{l=1}^{\frac{N}{2}} P(l)} \qquad (1)$$

where $P(l)$ is the power of the lth frequency bin, f_s is the sampling frequency, and N is the number of FFT points. The

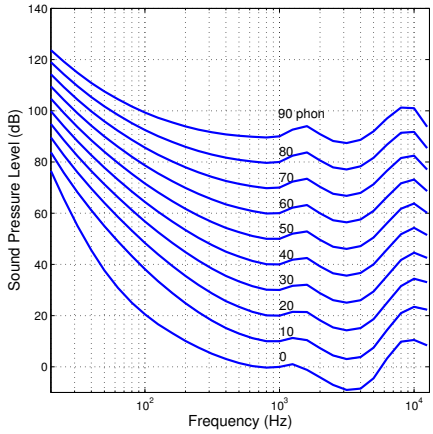

Fig. 2. Equal-loudness contour, where 'phon' is the unit of loudness.

source band is centered around M_1. The source band and the target band are of the same bandwidth, i.e., no frequency compression is applied in our method.

Weighted transposition: Transposing higher-frequency components to a lower-frequency band will result in some distortion. To reduce the distortion in the hearing experience, we propose to incorporate the concept of equal-loudness contour in our algorithm. Equal loudness is a psychoacoustics concept that represents the fact that the human auditory system has different sensitivity to different frequencies [14]. Fig. 2 depicts the equal-loudness contour according to the ISO standard in 2003 [15] based on a large number of human experiments conducted over the past decades. In our algorithm, we use the equal-loudness contour to calculate the difference in the sound pressure level (SPL) between the source frequency and the target frequency, and assign a set of weights to compensate this difference. The weights are given by $w(l) = 10^{(\frac{S_s(l)-S_t(l)}{20})}$, where $S_s(l)$ and $S_t(l)$ are the SPL of the lth frequency bin in source and target bands, respectively. Since the actual SPL of each frequency is unknown in the testing speech, we refer to the equal-loudness contour at 65 phon, which corresponds to the normal speaking level. The frequency-domain signal after transposition can be expressed as a weighted superposition of the original signal X_o and the source band signal X_s (with proper index shifting), i.e.,

$$X(l) = \begin{cases} X_o(l) + w(l-6)X_s(l-6), \\ \qquad \text{if } 750 < l \times (f_s/N) \le 1500\,. \\ X_o(l), \quad \text{if } l \times (f_s/N) \le 750 \end{cases} \qquad (2)$$

Signals below 750 Hz are left unchanged and signals above 1500 Hz are removed.

2.2. Evaluation Method

The Mandarin Monosyllable Recognition Test (MMRT) [16, 17] was conducted in a quiet meeting room for 8 na-

Table 1. Mandarin Consonants (Except /m/, /n/, /l/, /r/) by Places and Manners of Articulation

Place\Manner	Plosive	Affricate	Fricative
Labial	ㄅ/b/, ㄆ/p/		
Labiodental			ㄈ/f/
Front part of tongue tip		ㄗ/z/, ㄘ/c/	ㄙ/s/
Tongue tip	ㄉ/d/, ㄊ/t/		
Retroflex		ㄓ/zh/, ㄔ/ch/	ㄕ/sh/
Alveolar		ㄐ/j/, ㄑ/q/	ㄒ/x/
Velar	ㄍ/g/, ㄎ/k/		ㄏ/h/

tive Mandarin normal hearing (NH) listeners recruited from the Academia Sinica community. Testing on NH listeners precludes complicating factors associated with hearing loss (e.g., the duration and onset of hearing loss and hearing cell/auditory nerve survival rate) and hearing aid use (e.g., types of hearing aids used and signal processing algorithms), and allows us to explore the core ideas of this work. Our method is compared to the baseline method which implements a low-pass filter (LPF) with cut-off frequency of 1500 Hz. The LPF has been used to simulate hearing loss [18, 19]. Each subject listened to four lists of 25 phoneme-balanced Mandarin syllables processed by either the proposed method or the baseline method (i.e., each listener will listen to 100 FT-processed syllables and 100 LPF-processed syllables, which are randomly presented). All speech was normalized to have the same root mean squared (RMS) level and played by a SONY loudspeaker calibrated in 65 dB SPL in front of the listeners.

3. RESULTS AND DISCUSSION

We focus our discussion on Mandarin consonants except ㄇ/m/, ㄋ/n/, ㄌ/l/, ㄖ/r/ since they are sonorants and voiced consonants with formants similar to those of the vowels. The consonants under examination can be classified according to their places and manners of articulation, as summarized in Table 1.

3.1. Spectrogram

Fig. 3 depicts the spectrograms of the Mandarin syllable /qu/ (Tone 4) for the original, LPF-processed, and FT-processed speech. A visual examination of these spectrograms reveals that some high-frequency information is preserved in the lower-frequency region in the FT-processed speech as compared to the LPF-processed one.

3.2. Average Correct Identification of Consonants

Fig. 4 shows the average percentage of correct identification of consonants for LPF and FT methods. FT yields an 8% improvement overall, with most significant improvements in affricates and fricatives. Paired t-tests show statistically significant improvements overall ($p = 0.01$) and for fricatives

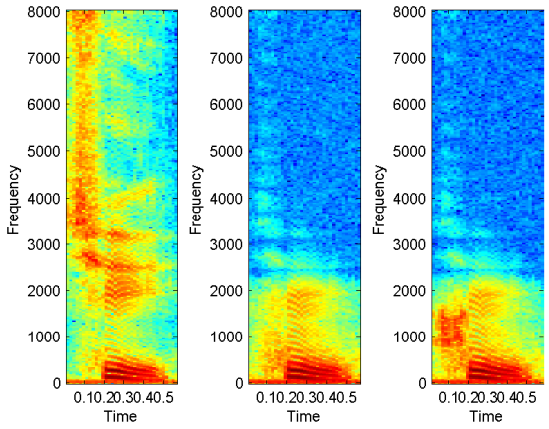

Fig. 3. Spectrograms of the Mandarin syllable /qu/ (Tone 4) for the original (left), LPF-processed (middle), and FT-processed speech (right).

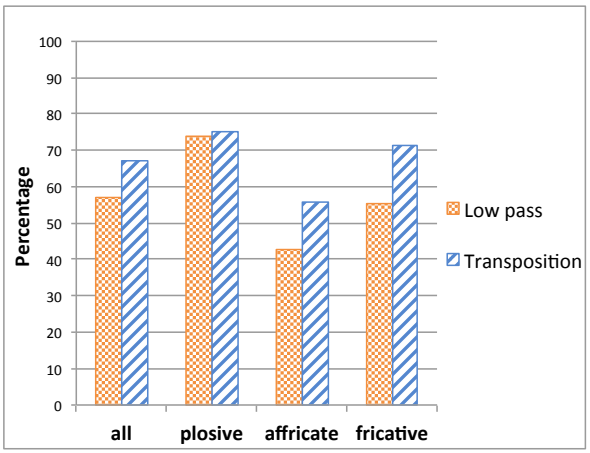

Fig. 4. Percentage of correct identification of Mandarin consonants (overall and by manners of articulation) for the baseline LPF and the proposed FT methods.

($p = 0.024$). The improvements for plosives ($p = 0.858$) and affricates ($p = 0.176$) are not significant. This may be explained by the large variation in performance across different subjects. For example, for affricates, significant improvements by FT were observed for six subjects but nearly no improvement for two subjects.

3.3. Average Correct Identification of Places and Manners of Articulation

Consonants contain manners- and places-of-articulation information corresponding to how the consonants are pronounced in the oral cavity. Fig. 5 exhibits the manners-of-articulation identification performance of the proposed FT scheme. If a consonant is misidentified as another consonant, while the correct and misidentified consonants are in the

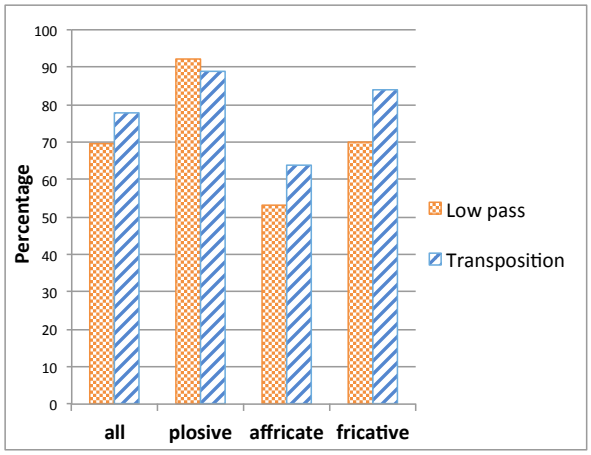

Fig. 5. Percentage of correct identification of manners of articulation of Mandarin consonants (overall and by manners of articulation) for the baseline LPF and the proposed FT methods.

same manner group (e.g., /z/ identified as /j/ in the affricate group), it is still considered a correct manners-of-articulation identification. From Fig. 5, a significant 7% improvement is obtained in overall identification with FT ($p = 0.041$), particularly contributed by improved affricates and fricatives identification. The FT scheme does not exhibit an advantage in plosives identification. The same investigation is conducted for places-of-articulation identification. The results show that a significant 6% improvement in overall identification is achieved with FT ($p = 0.04$; 66.54% for LPF and 73.06% for FT).

3.4. Consonant Confusions

The confusion matrix in Fig. 6 shows the percent change in confusion (FT subtracted by LPF, after rounding). The diagonals correspond to the correct identification of each syllable. We observe improved identification of syllables /b/,/d/, /j/, /q/, /x/, /zh/, /ch/, /sh/, /s/ by FT, especially /j/, /x/, and /sh/ with remarkable 30%, 38%, and 28% improvements, respectively. The "miss" in the confusion matrix records the percentage of neglecting the existence of a leading consonant and responding only with the vowel that follows the consonant. FT yields substantially lower miss rates. FT is ineffective for plosives such as /p/, /t/, /g/, /k/, with 16% more confusion from /p/ to /t/. The ineffectiveness of FT for plosives may be explained as follows. First, plosives are relatively low-frequency consonants and much shorter in length as compared to affricates and fricatives, and thus frequency transposition provides little benefit. Second, the percent correct identification of plosives is already as high as 90% for the LPF method (Fig. 5), suggesting that the additional high-frequency information can provide marginal advantages and thus the drawbacks of transposition (e.g., the masking of useful low-frequency information) outweigh its benefits. In Mandarin, retroflex consonants (e.g., /zh/, /ch/, /sh/) and their non-retroflex counterparts ("front part of tongue tip" in Table 1; e.g., /z/, /c/, /s/) are easily confused pairs. In fact, a great portion of incorrect identification of manners among affricates and fricatives (Fig. 5) is due to misidentifying /z/, /c/, /s/ as /zh/, /ch/, /sh/, respectively (about 33%). The confusion matrix reveals that the proposed FT can reduce the confusion by 4% from /z/ to /zh/, and 6% from /s/ to /sh/. However, there is 11% more confusion from /zh/ to /z/ in FT. The results suggest that the proposed FT scheme may enhance the discriminability in some syllables, although its potential for universal improvement in discriminability requires further study.

		Response																	
		/b/	/p/	/f/	/d/	/t/	/g/	/k/	/h/	/j/	/q/	/x/	/zh/	/ch/	/sh/	/z/	/c/	/s/	miss
Stimulus	ㄅ/b/	16	3	-9	-19					3		3						3	
	ㄆ/p/	-3	-6	-3	-3	16			-3	9	-3	3		3					-9
	ㄈ/f/	3														-3			
	ㄉ/d/				6														-9
	ㄊ/t/		-3			-13	3	3		6				3					
	ㄍ/g/	3					-9	3											3
	ㄎ/k/					3		-9						3	3				
	ㄏ/h/													3			-3		
	ㄐ/j/	-5		3		-3				30		-5	-6		3				-17
	ㄑ/q/									-8	13				-4				
	ㄒ/x/	-3			-5		3			-15	5	38			3				-23
	ㄓ/zh/				-9	2	4						4	-2	2	11			-5
	ㄔ/ch/		6	-9		-25			16					13					
	ㄕ/sh/			-3	-9	-9	-3		3					3	28		3	-3	-6
	ㄗ/z/				4		4						-4		-4	-4		-8	-4
	ㄘ/c/		25					-13											-13
	ㄙ/s/			-6	6							6	3	-13	-6			9	

Fig. 6. The confusion matrix shows the percent change in confusion (FT subtracted by LPF, after rounding), where blank entries indicate either the same percentage of confusion or no confusion by both LPF and FT.

4. CONCLUSION

We have proposed a new frequency transposition algorithm for Mandarin-speaking hearing aid users. The proposed method demonstrates an average 8% improvement on consonant recognition as compared to the LPF method in a simulated severe high-frequency hearing loss condition (1500 Hz cut-off frequency). A comparison among consonants of different manners of articulation reveals more significant improvements for affricates and fricatives than plosives. Improving the performance potential of the proposed algorithm in Mandarin consonant recognition (especially plosives) and conducting tests on hearing-aid users are worthwhile future work.

5. REFERENCES

[1] C. A. Hogan and C. W. Turner, "High-frequency audibility: Benefits for hearing-impaired listeners," *The Journal of the Acoustical Society of America*, vol. 104, no. 1, pp. 432–441, 1998.

[2] C. W. Turner and K. J. Cummings, "Speech audibility for listeners with high-frequency hearing loss," *American Journal of Audiology*, vol. 8, no. 1, pp. 47–56, 1999.

[3] A. Simpson, "Frequency-lowering devices for managing high-frequency hearing loss: A review," *Trends in Amplification*, vol. 13, no. 2, pp. 87–106, 2009.

[4] M. P. Posen, C. M. Reed, and L. D. Braida, "Intelligibility of frequency-lowered speech produced by a channel vocoder," *Journal of rehabilitation research and development*, vol. 30, pp. 26–26, 1993.

[5] Y.-Y. Kong and A. Mullangi, "On the development of a frequency-lowering system that enhances place-of-articulation perception," *Speech communication*, vol. 54, no. 1, pp. 147–160, 2012.

[6] A. Simpson, A. A. Hersbach, and H. J. McDermott, "Improvements in speech perception with an experimental nonlinear frequency compression hearing device," *International Journal of Audiology*, vol. 44, no. 5, pp. 281–292, 2005.

[7] D. Glista, S. Scollie, M. Bagatto, R. Seewald, V. Parsa, and A. Johnson, "Evaluation of nonlinear frequency compression: Clinical outcomes," *International Journal of Audiology*, vol. 48, no. 9, pp. 632–644, 2009.

[8] F. Kuk, D. Keenan, P. Korhonen, and C.-c. Lau, "Efficacy of linear frequency transposition on consonant identification in quiet and in noise," *American Academy of Audiology*, vol. 20, no. 8, pp. 465–479, 2009.

[9] J. D. Robinson, T. Baer, and B. C. Moore, "Using transposition to improve consonant discrimination and detection for listeners with severe high-frequency hearing loss," *International Journal of Audiology*, vol. 46, no. 6, pp. 293–308, 2007.

[10] Y.-T. Liu, Y. Tsao, and R. Y. Chang, "A deep neural network based approach to Mandarin consonant/vowel separation," in *IEEE International Conference on Consumer Electronics - Taiwan (ICCE- TW)*, June 2015.

[11] M. Velmans, "Speech imitation in simulated deafness, using visual cues and recoded auditory information," *Language and Speech*, vol. 16, no. 3, pp. 224–236, 1973.

[12] A. Jongman, R. Wayland, and S. Wong, "Acoustic characteristics of English fricatives," *The Journal of the Acoustical Society of America*, vol. 108, no. 3, pp. 1252–1263, 2000.

[13] K. Maniwa, A. Jongman, and T. Wade, "Acoustic characteristics of clearly spoken English fricatives," *The Journal of the Acoustical Society of America*, vol. 125, no. 6, pp. 3962–3973, 2009.

[14] H. Fletcher and W. Munson, "Loudness, its definition, measurement and calculation," *Journal of the Acoustic Society of America*, vol. 5, pp. 82–108, 1933.

[15] ISO226:2003, "Normal equalloudness level contours," *International Organization for Standardization*, 2003.

[16] K.-S. Tsai, L.-H. Tseng, C.-J. Wu, and S.-T. Young, "Development of a Mandarin monosyllable recognition test," *Ear and Hearing*, vol. 30, no. 1, pp. 90–99, 2009.

[17] PAD-MMRT speech recognition software. [Online]. Available: http://www.citi.sinica.edu.tw/papers/yu.tsao/4596-F.pdf

[18] H. McDermott and M. Dean, "Speech perception with steeply sloping hearing loss: effects of frequency transposition," *British Journal of Audiology*, vol. 34, no. 6, pp. 353–361, 2000.

[19] C. Füllgrabe, T. Baer, and B. C. Moore, "Effect of linear and warped spectral transposition on consonant identification by normal-hearing listeners with a simulated dead region," *International Journal of Audiology*, vol. 39, no. 6, pp. 420–433, 2010.