NONNEGATIVE MATRIX FACTORIZATION-BASED FREQUENCY LOWERING TECHNOLOGY FOR MANDARIN-SPEAKING HEARING AID USERS

Yen-Teh Liu, Yu Tsao, and Ronald Y. Chang*

Research Center for Information Technology Innovation, Academia Sinica, Taipei, Taiwan
*Email: yu.tsao@citi.sinica.edu.tw

ABSTRACT

Frequency lowering technologies have demonstrated effectiveness in English speech recognition for English-speaking people with high-frequency hearing loss. Their effect on Mandarin speech has not been well investigated. This paper serves two important purposes: it 1) examines the effect of frequency transposition (FT), a category of frequency lowering technologies, on Mandarin speech recognition, and 2) proposes a dictionary-based FT framework based on nonnegative matrix factorization (NMF) that is transferable across languages. Our results show that the proposed NMF-FT improves Mandarin consonant identification as compared to the traditional FT, with particularly significant improvements in affricates and fricatives.

Index Terms— Frequency lowering technology, Mandarin speech recognition, nonnegative matrix factorization (NMF).

1. INTRODUCTION

High-frequency hearing loss is a common type of hearing loss. A predominant cause of it is senile degeneration. According to World Health Organization (WHO), one-third of the world's population aged 65 or above suffer from different degrees of hearing loss [1]. Uncorrected hearing loss may associate with loneliness, withdrawal from social activities, sense of exclusion, etc., leading to degraded quality of life. Traditional amplifying hearing aids can benefit people of mild-to-moderate high-frequency hearing loss but provide limited improvements for those with severe-to-profound high-frequency hearing loss [2].

Frequency lowering has been proposed as a promising approach to combating high-frequency hearing loss [3]. This includes various signal processing techniques such as channel vocoder, frequency compression, and frequency transposition (FT), all presenting high-frequency information in a lower-frequency region accessible for people with high-frequency hearing loss. *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]. *Frequency transposition (FT)* shifts the high-frequency components to a lower-frequency band and adds them to the unprocessed lower-frequency signals [7, 8]. The FT method became the first frequency-lowering technique implemented in a commercial hearing aid [3].

The existing methods of FT have focused on the English language by incorporating language-specific characteristics of the language which may not represent other languages (e.g., Mandarin). We are motivated to propose a dictionary-based FT framework based on nonnegative matrix factorization (NMF) which can be easily adapted to different languages by substituting in proper dictionaries in a specific language. NMF can learn the basis matrix effectively with a small amount of training data and has found many speech applications, e.g., speech enhancement by utilizing small number of noise samples to train a noise basis matrix and effectively recover speech in noise [9], voice separation [10] and voice conversion [11] with limited training data. Our work inherits the core ideas of source separation and voice conversion, and presents a novel version of frequency lowering technology that demonstrates improved Mandarin speech recognition in simulated high-frequency hearing loss conditions.

This paper is organized as follows. Sec. 2 describes the proposed methods and the testing procedure. Sec. 3 presents the testing results and discussion. Sec. 4 concludes the paper and outlines the future work.

2. METHODS

2.1. Nonnegative matrix factorization

The NMF technique projects the columns of a nonnegative matrix $\mathbf{V}$ onto a space spanned by the basis vectors in $\mathbf{W}$ such that

$$\mathbf{V} \approx \mathbf{WH} \tag{1}$$

where $\mathbf{V} \in \mathbb{R}_+^{n\times m}$, $\mathbf{W} \in \mathbb{R}_+^{n\times r}$, and $\mathbf{H} \in \mathbb{R}_+^{r\times m}$ are all nonnegative matrices ($\mathbb{R}_+$ denotes the set of nonnegative real numbers), with r often chosen to be smaller than m. Matrix $\mathbf{W}$ is often called the *basis matrix* and matrix $\mathbf{H}$ the *coefficient matrix* of $\mathbf{V}$. The factorization in (1) is commonly

approximated by finding $\mathbf{W}$ and $\mathbf{H}$ that minimize

$$D = \|\mathbf{V} - \mathbf{W}\mathbf{H}\|_F^2 \tag{2}$$

where $\|\cdot\|_F$ denotes the Frobenius norm. $\mathbf{W}$ and $\mathbf{H}$ can be obtained by the following algorithm that alternately updates the matrices from iteration i to iteration $i+1$ [12]:

$$\mathbf{W}_{i+1} = \mathbf{W}_i \cdot \frac{\mathbf{V}\mathbf{H}_i^T}{\mathbf{W}_i\mathbf{H}_i\mathbf{H}_i^T} \tag{3}$$

$$\mathbf{H}_{i+1} = \mathbf{H}_i \cdot \frac{\mathbf{W}_{i+1}^T\mathbf{V}}{\mathbf{W}_{i+1}^T\mathbf{W}_{i+1}\mathbf{H}_i} \tag{4}$$

where $(\cdot)^T$ denotes the transpose operation.

2.2. NMF-based frequency transposition (NMF-FT)

Mandarin consonants and vowels have different spectral characteristics. Consonants comprise higher-frequency components and vowels have significant formant structures. The basic idea of our approach is to use NMF to train $\mathbf{W}$ (hereafter referred to as *dictionary*) that characterizes both consonants and vowels, and then replace the consonant dictionary by one that incorporates some frequency lowering processing. We can easily adapt to another language other than Mandarin through training another set of dictionaries in that language. Compared to traditional frequency lowering methods that compress or transpose the high-frequency speech into low-frequency bands, the proposed method has several advantages: 1) frequency lowering is performed in a dictionary-based manner and the dictionaries can be easily customized to meet specific requirements; 2) frequency lowering is performed on the segment level rather than the frame level and thus can alleviate possible discontinuity issues. The proposed method is comprised of a training stage and a transposition stage. In the training stage, three dictionaries $\mathbf{W}_c$, $\mathbf{W}_v$, and $\mathbf{W}_{ct}$ are prepared based on the training data sets of consonants $\mathbf{V}_c$, vowels $\mathbf{V}_v$, and frequency-transposed consonants $\mathbf{V}_{ct}$, respectively. All training data are recorded by a male speaker with 16 kHz sampling frequency. A Hamming window of length 128 and a 128-point FFT are applied on the speech signals to construct the spectrograms for NMF. The transposition stage is an online processing where the original testing speech is transposed to a frequency-lowered one based on the three dictionaries from the training stage. The proposed method is described in detail below.

2.2.1. Training stage

Dictionaries $\mathbf{W}_c$ and $\mathbf{W}_v$ are trained according to

$$\mathbf{V}_c \approx \mathbf{W}_c\mathbf{H}_c, \quad \mathbf{V}_v \approx \mathbf{W}_v\mathbf{H}_v. \tag{5}$$

In order to train the dictionary of frequency-lowered consonants, $\mathbf{W}_{ct}$, we prepare a set of training data $\mathbf{V}_{ct}$ which is obtained by performing a high-frequency to low-frequency transposition on $\mathbf{V}_c$. Since this stage is performed offline, the transposition can be carefully conducted. In this study, we consider a severe hearing loss condition where signal components above 1500 Hz are assumed inaudible. To capture the most relevant high-frequency spectral components on each phoneme, we transpose the high-frequency spectral components based on the first spectral moment M_1 calculated for each phoneme above 1500 Hz, where M_1 represents the centroid of the spectrum, to characterize the consonants [13, 14]. M_1 is calculated as

$$M_1 = \frac{\sum_{l=13}^{\frac{N}{2}} P(l)(f_s \times l/N)}{\sum_{l=13}^{\frac{N}{2}} P(l)} \tag{6}$$

where $P(l)$ is the power of the lth frequency bin ($l = 13$ corresponds to the starting frequency bin at 1500 Hz), f_s is the sampling frequency, and $N = 128$ is the number of FFT points. The frequency bins around M_1 are transposed to lower frequency band only.

After obtaining $\mathbf{V}_{ct}$, a direct estimation of $\mathbf{W}_{ct}$ can be carried out by following (2). However, since NMF does not impose constraints on the order of the basis vectors when updating matrix parameters, the order of basis vector in the directly estimated $\mathbf{W}_{ct}$ may not align with that in $\mathbf{W}_c$. The misalignment can cause distortion in the transposition stage. To address this issue, we adopt the NMF adaptation technique [15] to estimate $\mathbf{W}_{ct}$. The NMF adaptation technique introduces an additional matrix $\mathbf{A}$ into the original NMF framework, and the frequency-lowered dictionary is estimated by $\mathbf{W}_{ct} = \mathbf{A}\mathbf{W}_c$. Similar to (2), to compute $\mathbf{A}$, we define

$$D_a = \|\mathbf{V}_{ct} - \mathbf{A}\mathbf{W}_c\mathbf{H}\|_F^2 . \tag{7}$$

In this study, we choose matrix $\mathbf{A}$ as a diagonal matrix. In the training stage, we first set $\mathbf{A}$ to an identity matrix, and $\mathbf{H}$ to $\mathbf{H}_c$. Then, we update $\mathbf{A}$ and $\mathbf{H}$ alternately according to:

$$\mathbf{A}_{i+1} = \mathbf{A}_i \cdot \frac{\mathbf{V}_{ct}(\mathbf{W}_c\mathbf{H}_i)^T}{\mathbf{A}_i(\mathbf{W}_c\mathbf{H}_i)(\mathbf{W}_c\mathbf{H}_i)^T} \tag{8}$$

$$\mathbf{H}_{i+1} = \mathbf{H}_i \cdot \frac{(\mathbf{A}_{i+1}\mathbf{W}_c)^T\mathbf{V}_{ct}}{(\mathbf{A}_{i+1}\mathbf{W}_c)(\mathbf{A}_{i+1}\mathbf{W}_c)^T\mathbf{H}_i}. \tag{9}$$

Note that $\mathbf{W}_c$ is fixed in the above update procedure.

2.2.2. Transposition stage

In the transposition stage, $\mathbf{W}_c$ and $\mathbf{W}_v$ are continued to be used to provide suitable dictionaries in the testing speech. The two matrices are horizontally concatenated in the form of $\mathbf{W}_T = [\mathbf{W}_c\ \mathbf{W}_v]$, where $\mathbf{W}_T \in \mathbb{R}_+^{n\times 2r}$ is the dictionary of the testing data matrix $\mathbf{V}_T$ in the NMF process. The testing data matrix $\mathbf{V}_T$ is composed of the magnitude spectrogram of the testing speech and can be approximated by NMF:

$$\mathbf{V}_T \approx \mathbf{W}_T\mathbf{H}_T = [\mathbf{W}_c\ \mathbf{W}_v]\begin{bmatrix}\mathbf{H}_c\\ \mathbf{H}_v\end{bmatrix}. \tag{10}$$

Different from the training stage which alternately updates $\mathbf{W}$ and $\mathbf{H}$, here dictionary $\mathbf{W}_T$ remains unchanged while the coefficient matrix $\mathbf{H}_T$ keeps updating iteratively to achieve a better factorization approximation.

Once satisfactory approximation is achieved, we replace $\mathbf{W}_c$ by the frequency-transposed consonant dictionary $\mathbf{W}_{ct}$ to form a new dictionary $\mathbf{W}'_T = [\mathbf{W}_{ct}\ \mathbf{W}_v]$. The frequency lowering NMF-FT magnitude spectrogram for testing speech hence constitutes the new dictionary $\mathbf{W}'_T$, i.e.,

$$\mathbf{V}'_T \approx \mathbf{W}'_T\mathbf{H}_T = [\mathbf{W}_{ct}\ \mathbf{W}_v]\begin{bmatrix}\mathbf{H}_c\\ \mathbf{H}_v\end{bmatrix}. \qquad (11)$$

Finally, the frequency-lowering spectrogram $\mathbf{V}'_T$ together with the original phase components are converted to the time domain via IFFT to produce the frequency lowering speech.

2.3. Test procedure

The Mandarin Monosyllable Recognition Test (MMRT) [16] was conducted in a quiet meeting room for 8 native 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. The low-pass filter (LPF) with a cut-off frequency of 1500 Hz is used to simulate hearing loss, similar to [17, 18]. Our NMF-FT method is compared to two baseline methods: the LPF only scheme, and the traditional FT scheme (transposition from 1500–3000 Hz to the frequency band below 1500 Hz after LPF). Each subject listened to four lists of 25 phoneme-balanced Mandarin syllables processed by one of the three methods under test (i.e., each listener will listen to 100 NMF-FT-processed syllables, 100 LPF-processed syllables, and 100 FT-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 these 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. Average correct identification of consonants

Fig. 1 shows the average percentage of correct identification of consonants for LPF, FT, and NMF-FT methods. FT yields a 4% overall improvement over LPF. The improvement is however not statistically significant as shown by one-way ANOVA within subjects ($p = 0.398$). NMF-FT exhibits a 12% overall improvement over LPF. Comparing NMF-FT and LPF, statistically significant improvements overall ($p = 0.017$) and for affricates ($p = 0.015$) and fricatives ($p = 0.007$) are observed. Notably, NMF-FT improves the identification of fricatives by about 14% while FT achieves little improvement, as compared to LPF. NMF-FT's improvement for plosives is not significant as compared to LPF ($p = 0.307$), which may be explained by the large variation in performance across different subjects.

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/

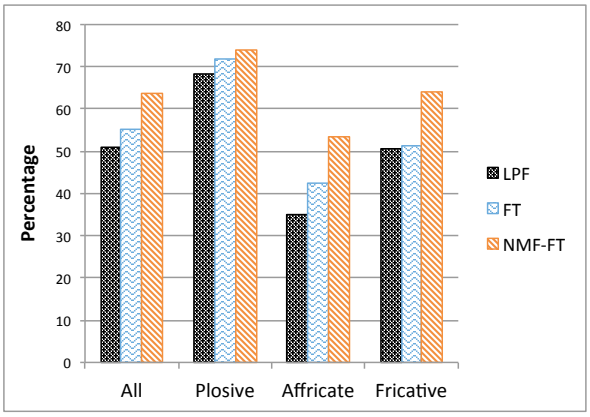

Fig. 1. Percentage of correct identification of Mandarin consonants (overall and by manners of articulation) for the baseline LP and FT and the proposed NMF-FT methods.

3.2. Average correct identification of places and manners of articulation

Consonants contain places- and manners-of-articulation information corresponding to how the consonants are pronounced in the oral cavity. Fig. 2 shows the manners-of-articulation performance of the proposed NMF-FT scheme. If a consonant is misidentified as another consonant in the same manner group (e.g., /j/ identified as /zh/ in the affricate group), it is considered correct manners-of-articulation identification. From Fig. 2, significant 18% and 9% improvements

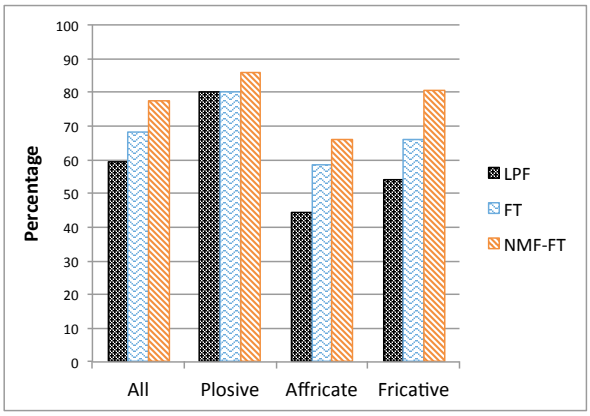

Fig. 2. Percentage of correct identification of manners of articulation of Mandarin consonants (overall and by manners of articulation) for the baseline LPF and FT and the proposed NMF-FT methods.

		Response																	
		/b/	/p/	/f/	/d/	/t/	/g/	/k/	/h/	/j/	/q/	/x/	/zh/	/ch/	/sh/	/z/	/c/	/s/	miss
Stimulus	ㄅ/b/	**12**		-19	12					-7									
	ㄆ/p/		**3**		-7	6			3	3									-10
	ㄈ/f/	-10		**6**			-4		-4					6			3		
	ㄉ/d/	3			**9**	-4	-4			-4						3			-7
	ㄊ/t/		3	-4		**9**		-10	3										-4
	ㄍ/g/				-4		**3**						3						-4
	ㄎ/k/							**-4**						3					
	ㄏ/h/							-7	**-38**			9	3	6					25
	ㄐ/j/	-4			-11	3				**37**						-2	-2		-25
	ㄑ/q/				-9				4	-13	**12**				4		4		-5
	ㄒ/x/				-7				4	-15	6	**62**		2					-42
	ㄓ/zh/			2	-3		-3			2			**8**	-5		6		-7	-5
	ㄔ/ch/			-4		-19			-7					**15**			12		
	ㄕ/sh/	-4		6	-16			6	9				-10	-4	**9**	-10	3	12	-7
	ㄗ/z/			-7		6	12						6			**12**		-25	
	ㄘ/c/		-7	6		-25	0	6	6					18			**-7**		
	ㄙ/s/			12	-4				3				-16		15	-13		**3**	

Fig. 3. Confusion matrix showing the percent change in confusion (NMF-FT subtracted by LPF, after rounding), where blank entries indicate either the same percentage of confusion or no confusion by both LPF and NMF-FT.

are achieved for NMF-FT in overall identification over LPF and FT, respectively ($p = 0$ and $p = 0.034$, respectively), especially contributed by improved affricates and fricatives identification. Both FT and NMF-FT do not exhibit statistically significant improvements in plosives identification as compared to LPF. The higher percentage increases in Fig. 2 as compared to Fig. 1 (by counting in misidentification in the same manner group) for NMF-FT suggest that misidentification in NMF-FT more likely occurs in the same manner group as compared to LPF. The same evaluation is conducted for places-of-articulation identification. The results show that there is no statistically significant improvement in places-of-articulation identification for NMF-FT as compared to LPF ($p = 0.132$).

3.3. Consonant confusions

The confusion matrix in Fig. 3 shows the percent change in confusion (NMF-FT subtracted by LPF, after rounding). The diagonals correspond to correct identification of each syllable. We observe a general trend of improvement by NMF-FT, including syllables /b/, /p/, /f/, /d/, /t/, /g/, /j/, /q/, /x/, /zh/, /ch/, /sh/, /z/, /s/, with remarkable advantages on /j/ and /x/ (37% and 62%, respectively). However, NMF-FT yields 38% degraded performance on syllable /h/. The improvement on /zh/, /ch/, /sh/, /z/, /s/ is not as significant as on /j/ and /x/ and there is a slight degradation on /c/ for NMF-FT. This may be attributed to the fact that, in Mandarin, retroflex consonants (/zh/, /ch/, /sh/) and their non-retroflex counterparts ("front part of tongue tip" in Table 1; /z/, /c/, /s/) are easily confused pairs. In fact, a great portion of incorrect identification of manners among affricates and fricatives (Fig. 2) is due to misidentification between these pairs. Fig. 3 reveals that there is more confusion in NMF-FT from /zh/, /ch/, /sh/ to /z/, /c/, /s/, respectively (6%, 12%, 12%, respectively), and from /z/, /c/, /s/ to /zh/, /ch/, /sh/, respectively (6%, 18%, 15%, 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. NMF-FT yields lower miss rates in general.

4. CONCLUSION

We have proposed a new frequency transposition framework based on NMF. The proposed method is transferable across different languages. The proposed NMF-FT method demonstrates 12% and 8% overall improvements on Mandarin consonant recognition as compared to LPF and traditional FT methods, respectively, in a simulated high-frequency hearing loss condition. The proposed method yields more significant improvements on affricates and fricatives than plosives. The results suggest that the proposed method improves the discriminability in most Mandarin syllables but its potential for universal improvement requires further study. Testing the proposed method on different languages to leverage the proposed method's easy adaptation to different languages is also a worthwhile future work.

5. REFERENCES

[1] World Health Organization, (2012)WHO global estimates on prevalence of hearing loss. Available: http://www.who.int/pbd/deafness/estimates/en/

[2] C. A. Hogan and C. W. Turner, "High-frequency audibil-

ity: Benefits for hearing-impaired listeners", *The Journal of the Acoustical Society of America*, vol. 104, no. 1, pp.432-441, 1998.

[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] 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.

[9] K. W. Wilson, B. Raj, P. Smaragdis, and A. Divakaran, "Speech denoising using nonnegative matrix factorization with priors", *in Proc. International Conference on Acoustics, Speech and Signal Processing*, pp. 4029–4032, 2008.

[10] M. Schmidt and R. Olsson, "Single-channel speech separation using sparse non-negative matrix factorization", *in Proc. Interspeech*, 2006.

[11] R. Takashima, R. Aihara, T. Takiguchi, and Y. Ariki, "Noise-robust voice conversion based on spectral mapping on sparse space", *in Proc. International Speech Communication Association*, pp. 71–75, 2013.

[12] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization", *in Proc. Advances in Neural Information Processing Systems*, pp. 556–562, 2001.

[13] 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.

[14] 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.

[15] E. M. Grais and H. Erdogan, "Adaptation of speaker-specific bases in non-negative matrix factorization for single channel speech-music separation", *in Proc. International Speech Communication Association*, 2011.

[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] 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.

[18] C. Fullgrabe, 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.