PERCEIVED QUALITY OF RESONANCE BASED DECOMPOSED SPEECH COMPONENTS UNDER DIOTIC AND DICHOTIC LISTENING

Chin-Tuan Tan[1], Ivan W. Selesnick[2] and Kemal Avci[3]

[1]New York University, School of Medicine, Department of Otolaryngology, USA
[2]Polytechnic Institute of New York University, Department of Electrical & Computer Engineering, USA
[3]Abant Izzet Baysal University, Department of Electrical & Electronics Engineering, Turkey
E-mail: chin-tuan.tan@nyumc.org, selesi@poly.edu and avci_k@ibu.edu.tr

ABSTRACT

This study investigates the feasibility of using binaural dichotic presentation of speech components decomposed using a recently proposed resonance-based decomposition method to release listeners from intra-speech masking and yield better perceived sound quality. Resonance-based decomposition is a nonlinear signal analysis method based not on frequency or scale but on resonance. We decomposed different categories of speech stimuli (vowels, consonants, and sentences) into low- and high-resonance component using various combination of low- and high-Q-factors {Q1,Q2}. 10 normal hearing listeners were asked to rate the perceived quality of each individual decomposed component presented diotically, and in pair presented dichotically. We found that the perceived quality rating of these resonance components when presented in pair was higher than the mean of perceived quality ratings of these resonance components when presented individually. Our result suggests that listeners were able to fuse binaural dichotic presentation of high- and low-resonance components and perceived better sound quality.

Index Terms— Resonance-based decomposition, dichotic representation of speech, binaural fusion

1. INTRODUCTION

Intra-speech masking refers to the interference in the perception of speech cues by components of the speech signal itself in either spectral or temporal domain. In normal adult speech produced at a typical conversational level of 63 dB SPL [17], the greatest acoustic energy is located in the region below 800 Hz and steadily decreases at higher frequencies, which matches the frequency ranges associated with first formant (F1) and second formant (F2) based on adult productions of ten English vowels [19]. A greater intensity of F1 relative to F2 is likely to mask F2. This process by which the audibility of an acoustic stimulus is reduced in the presence of acoustic energy at lower frequencies is known as upward spread of masking. Summers and Leek [23] showed that the consonant-vowel identification for both hearing impaired (HI) and normal hearing (NH) listeners was improved by increasing attenuation of the first formant (F1) up to 18 dB. This benefit associated with F1 attenuation was more consistently present for HI than for NH subjects with no difference in magnitude for testing in quiet and in noise, particularly at high signal presentation levels. They also found that the influence of the proximity of F1 and F2 in vowels may have outweighed the greater level difference between F1 and F2. But at high presentation levels, F1-on-F2 masking may occur for vowels with considerable frequency spacing between F1 and F2. Similarly, Danaher and Pickett [6] also attempted to reduce the F1 effect by presenting F1 and F2 to separate ears of a subject. NH listeners did not show the F1 on F2 masking effects reported for HI listeners at comfortable listening levels. However, both groups showed susceptibility to F1 on F2 masking at high-presentation levels. Van Tasell [26] also observed similar outcomes at a lower presentation level.

One reason for this might be that HI subjects often show a reduced frequency selectivity [7][18][23] compared to NH subjects and a high variability in specific binaural tasks. Their binaural abilities greatly decrease with increasing hearing loss [9][20][27]. However, most of these studies measured the frequency selectivity monaurally. In a recent study, Nitschmann et al. [15] showed that the auditory filter widths derived from data for signals with interaural disparities (dichotic condition) are usually larger than those derived from listening conditions where the same stimuli were presented to both ears (diotic condition) for both NH and HI listeners using the common notched-noise experiment [16] in a similar paradigm as used in [8]. They found that both NH and HI subjects show wider auditory filters in the dichotic notched-noise experiment. However, the ratios of binaural divided by monaural auditory filter

This work is supported in part by NIH/NIDCD 1K25DC010834-01 (PI: Tan) and NSF CCF-1018020 (PI: Selesnick)

bandwidths were similar between NH and HI subjects indicating that the binaural specific and centrally located processing has effectively increased the auditory filter width in dichotic listening situations for both NH and HI subjects. The ratios that were greater than 1.0 could be caused by larger monaural bandwidths at the peripheral stage of processing before the binaural synchrony of the input from both ears and lead to a larger binaural bandwidth.

Clearly reduction in the effects of intra-speech masking by additional signal processing like spectral contrast enhancement [2][5] and multi-band frequency compression [1][10] is more beneficial for listening speech in complex environments, particularly for HI listeners. However these techniques generally degrade speech quality and are difficult to implement using low-power signal processors in hearing aids. Hence, several studies have made use of binaural dichotic presentation by spectrally splitting the speech signal using a pair of comb filters with complementary magnitude responses[3][4][11][12][13]. But the differences in the bandwidths and the magnitude responses of the filters used in these studies yield great variation in the results reported.

In this study, we attempt to address the issue by splitting the speech signal into high- and low- resonance components using a resonance-based signal decomposition method [9]. We also make use of binaural dichotic presentation of these two components in attempt to release both normal hearing and hearing impaired listener from intra-speech masking without compromising the perceived sound quality [24]. In this paper, we examined the efficacy of this method with consonants, vowels, and sentences, and compared the perceived quality of these decomposed components when they are presented individually and in pairs.

In Section 2, we describe the resonance-based decomposition method. We then describe the subjective listening experiment in Section 3 and the results in Section 4. Finally, conclusion and future work are presented in the last section.

2. RESONANCE BASED DECOMPOSITION

Resonance-based signal decomposition method [21] is a nonlinear signal analysis method that decomposes a signal into high-resonance (oscillatory) and low-resonance (transient) components. It utilizes sparse signal representations, morphological component analysis, and constant-Q wavelet transforms [22] with adjustable Q factor. In this method, the speech waveform is modeled as

$$x = x_1 + x_2 + n \text{ with } x, x_1, x_2 \in \Re^N \text{ x, } x_1, x_2$$

where the components x_1 and x_2 are the high and low resonance components, and n is noise. More information for speech decomposition based on this method can be found in our previous work [24].

3. SUBJECTIVE LISTENING EXPERIMENT

Ten English-speaking NH listeners (five male and five female) between the ages of 18 and 32 were asked to rate the quality of resonance-based decomposed components as they were perceived. All listeners have normal hearing with pure tone hearing thresholds better than 20 dB HL between 500 and 6000 Hz in both ears.

Each speech stimulus used in the experiment was decomposed into resonance components with the following combinations of Q factors, i.e. {Q1,Q2} = {1,2; 1,3; 1,4; 1,5; 2,3; 2,4; 2,5; 3,4; 3,5; 4,5}. Under this combination, there is a total of 20 decomposed components with 10 low-resonance components, i.e. Q1 = {1 ; 1 ; 1 ; 1 ; 2 ; 2 ; 2 ; 3 ; 3 ; 4 }, and 10 high-resonance components, i.e. Q2 = { 2; 3; 4; 5; 3; 4; 5; 4; 5; 5}. For instance, the low-resonance component with Q1=1 decomposed with the combination {Q1=1,Q2=2} and that decomposed with the combination {Q1=1,Q2=3} are considered as two unique components. The resonance components were presented to the listeners individually with one of the 20 resonance components, or in pairs with the above combinations {Q1,Q2}. A diotic presentation of individual resonance component refers to either low-resonance component Q1 or high-resonance component Q2 was presented to both left and right ears simultaneously. Dichotic presentation of high- and low-resonance in the combination {Q1,Q2} refers to Q1 was presented to left ear while Q2 was presented to right ear simultaneously. In addition, Q1 was also presented to right ear and Q2 was presented to left ear for examining the lateralization effect. Two tests were conducted for each stimulus. Before starting each stimulus test, the original clean version of the stimulus was presented to the listener. The resonant components of the stimulus were then randomly presented to the listeners and they were asked to listen to the decomposed components and rate their perceived quality on a 10-point rating scale where '10' indicates 'clean – completely undistorted' and '1' represents 'very distorted'.

Altogether, three sets of speech stimuli were used in this listening experiment. The first set of consonant stimuli comprised of sixteen consonants, spoken by a male talker, in /aCa/ context (e.g., the consonant /b/ would be presented as /aba/). These stimuli were obtained from the Iowa Audiovisual Speech Perception Laser Video Disc [25]. The second set of vowel stimuli consisted of nine vowels, spoken by a female talker, in an /hVd/ context (e.g., the vowel /ea/ would be presented as /head/). The third set of stimuli consisted of ten sentences, spoken by a male talker, in Hearing in Noise Test (HINT) context [14].

The decomposed components of each speech stimulus were processed and presented to listeners using a Creative Sound Blaster digital audio card sound card, mounted in a PC. The output of the sound card drove Sennheiser HD580 headphones at a comfortable listening level.

4. RESULTS

All listeners reported that the resonance components of all stimuli were intelligible but their perceived quality varies with different combinations of high- and low-resonance components. They were able to complete the rating task with consonant, vowel and sentences stimuli, and their mean ratings across listeners separately tabulated for vowel, consonant and sentence stimuli in Table 1, when they listened to each individual resonance component, and Table 2, when they listened to each pair of resonance components.

Consonant									
Q1									
		1		2		3		4	
Q2	2	7.55	8.11						
	3	7.36	8.16	6.95	7.83				
	4	7.08	8.02	6.63	7.59	6.41	7.37		
	5	6.91	8.03	6.53	7.63	6.18	7.31	6.09	7.04
Vowel									
Q1									
		1		2		3		4	
Q2	2	7.96	8.41						
	3	7.74	8.56	7.46	8.54				
	4	7.52	8.55	7.37	8.32	7.42	8.25		
	5	7.33	8.49	7.34	8.41	7.3	8.27	7.3	8.17
Sentence									
Q1									
		1		2		3		4	
Q2	2	7.44	8.47						
	3	7.02	8.48	6.7	8.47				
	4	6.72	8.44	6.74	8.4	6.88	8.34		
	5	6.55	8.41	6.67	8.39	6.94	8.27	7.12	8.27

Table 1: Mean quality ratings of individual resonance components of speech stimuli under diotic listening condition. The left and right values in each entry are mean ratings with the low- {Q1} and high- {Q2} resonance components decomposed under the combination {Q1, Q2}.

In Table 1, the mean quality ratings of 10 listeners for listening to each of the 20 resonance components was computed for each category of speech stimuli and tabulated in the form of a matrix with each entry indicating the combination {Q1,Q2} of which the resonance components were decomposed. The value of Q2 increases from top row to bottom row, while the value of Q1 increases from left column to right column. The left value of each entry referred to the mean quality rating of the low- resonance component Q1 presented to both ears while the right value of the entry referred to the mean quality rating of high resonance component Q2 presented in the same manner. The same tabulation was repeated for consonant, vowel and sentence stimuli.

In general, the mean quality rating of high-resonance component was higher than that of the complementary low-resonance component for all three categories of speech stimuli within a 2-point range on a 10-point scale, when they are listened individually in a diotic presentation. Most of the mean quality ratings of both high- and low-resonance components decomposed from vowel stimuli were higher than those obtained from the corresponding resonance components decomposed from consonant and sentence stimuli.

Consonant									
Q1									
		1		2		3		4	
Q2	2	7.66	7.57						
	3	7.74	7.73	7.46	7.55				
	4	7.82	7.85	7.46	7.51	7.32	7.21		
	5	7.72	7.76	7.43	7.44	7.1	7.04	6.99	6.95
Vowel									
Q1									
		1		2		3		4	
Q2	2	8.33	8.21						
	3	8.39	8.3	8.08	8.02				
	4	8.32	8.32	8.02	7.96	7.73	7.71		
	5	8.22	8.2	7.89	7.73	7.69	7.77	7.73	7.61
Sentence									
Q1									
		1		2		3		4	
Q2	2	8.02	8.08						
	3	8.03	8.2	8.06	8.05				
	4	8.15	8.24	8.03	8.04	7.88	7.99		
	5	8.21	8.18	8.11	8.09	7.81	7.92	7.8	7.96

Table 2: Mean quality ratings of each pair of high- and low-resonance components of speech stimuli under dichotic listening condition. The left values in each entry referred to mean quality ratings obtained with {Q1} presented to the left ear and {Q2} presented to the right ear {Q1, Q2} while the right values referred to those obtained in a reverse order {Q2,Q1}.

In all categories of speech stimuli, the smallest difference between the mean quality ratings of high- and low-resonance components were obtained when they were decomposed under the combination of {Q1=1, Q2=2}. It was not surprising to yield high- and low- resonance

components of similar sound quality with both Q-factors of values close to 1 and a small difference between them. Clearly, resonance-based decomposition method was able to re-distribute the perceptually information in the original signal to the high- and low- resonance components with this combination of Q-factors, when they are individually presented.

In Table 2, the mean quality ratings of 10 listeners for listening to each pair of high- and low- resonance components decomposed from one of 10 possible combinations of {Q1, Q2} were computed for each category of speech stimuli and tabulated in a similar format as in Table 1. Each pair of high- and low- resonance components were presented to the listeners in a dichotic manner. Here, the left value of each entry referred to the mean quality rating obtained when the low- resonance component Q1 was presented to left ear and the high- resonance component Q2 was presented to right ear {Q1,Q2}, while the right value of the entry referred to the mean quality rating obtained when the components were presented in a reverse order {Q2,Q1}. Apparently, the order of the presentation for these two components barely affects the mean quality ratings of the listeners. The absolute difference between the left and right values of each entry of the matrices in Table 2 varies between 0.01 and 0.16 on a 10-point scale for the three categories of speech stimuli. The mean of the left and right values of each entry of the matrices in Table 2 was higher than that of corresponding entry in Table 1. The mean quality ratings were suggesting that the listeners were able to make use of binaural dichotic presentation of high- and low-resonance components and perceive a better sound quality of speech than listening to each component alone. Unlike in Table 1, the highest means of the left and right values were found in different entry location of the matrix for each categories of speech stimuli; {Q1=1, Q2=3} for vowels, {Q1=1, Q2=4} for consonants, and {Q1=1, Q2=5} for sentences. The result seems to suggest that the resonance components decomposed using a combination {Q1, Q2} would require a greater difference in values between Q1 and Q2 to achieve a better binaural fusion and yield higher perceived sound quality.

5. CONCLUSION

Result presented in this paper with 10 NH listeners positively supports the use of the resonance-based decomposition method in splitting speech into a pair of binaural dichotic resonance components for possible release of some intra-speech masking and yield a better perceived sound quality for both NH and HI listeners. Clearly, the dichotic presentation of high- and low-resonance components does not result in a lateralization of sound and there exists a combination of Q-factors that would yield an 'adequate' splitting of the speech signal for better binaural fusion. In addition, the perceived quality rating of each individual resonance component is comparable in value to that of a pair of components suggests that 'noise' introduced by the resonance-based decomposition method barely affects the binaural fusion process, which is an advantage particularly for HI listeners with narrow dynamic range

Currently we are extending our study to include the perception of the resonance components scaled with equal intensity and equal loudness binaurally by NH and HI listeners of different binaural frequency selectivity, and a more binaural specific assessment task like sound localization. Since the process involved is more centrally located, we will also examine at cortical/brainstem response evoked by resonance components in near future.

6. REFERENCES

[1] T. Arai T, K. Yasu, and N. Hodoshima, "Effective speech processing for various impaired listeners," *Proc 18th Int Congress Acoust.* Kyoto, Japan, II, pp. 1389 – 1392, 2004.

[2] T. Baer, B.C.J. Moore, and S. Gatehouse, "Spectral contrast enhancement of speech in noise for listeners with sensorineural hearing impairment: Effects on intelligibility, quality, and response times," *J Rehabil Res Dev,* 30, pp. 49 – 72, 1993.

[3] D.S. Chaudhari, and P.C. Pandey, "Critical band splitting of speech signal for reducing the effect of spectral masking in bilateral sensorineural hearing impairment," *Proc 5th Int Symp Signal Proc and Its Applications (ISSPA 1999)*, Brisbane, Australia, pp. 119 – 122, 1999.

[4] A.N. Cheeran, and P.C. Pandey, "Evaluation of speech processing schemes using binaural dichotic presentation to reduce the effect of masking in hearing hearing-impaired listeners," *Proc 18th Int Congress Acoustics* , Kyoto, Japan, II, pp. 1523 – 1526, 2004.

[5] I. Cohen, "Speech spectral modeling and spectral enhancement based on autoregressive conditional heteroscedasticity models," *Signal Processing*, 86, pp. 698 – 709, 2006.

[6] E. M. Danaher, and J. M. Pickett, "Some masking effects produced by low-frequency vowel formants in persons with sensorineural hearing loss," *J. Speech Hear. Res.* 18, pp. 261– 271, 1975.

[7] B.R. Glasberg, and B.C.J. Moore, "Auditory filter shapes in subjects with unilateral and bilateral cochlear impairments," *J Acoust Soc Am* , 76(2), pp.419–427, 1986.

[8] J.W. Hall, R.S. Tyler, and M.A. Fernandes, "Monaural and binaural auditory frequency resolution measured using bandlimited noise and notched-noise masking," *J Acoust Soc Am* , 73(3), 894–898, 1983.

[9] Holube I., *Experiments and models concerning psychoacoustics and speech perception in normal-hearing and hearing-impaired listeners*, PhD thesis, Georg-August-Universität, Göttingen, 1993.

[10] P.N. Kulkarni, P.C. Pandey, and D.S. Jangamashetti, "Multi-band frequency compression for sensorineural hearing impairment," *Proc 16th Int Conf Digital Signal Processing*, Santorini, Greece, paper S4P.1, 2009.

[11] T. Lunner, S. Arlinger, and J. Hellgren, "8-channel digital filter bank for hearing aid use: Preliminary results in monaural, diotic, and dichotic modes," *Scand Audiol Suppl*, 38, pp. 75 – 81, 1993.

[12] P.E. Lyregaard, "Frequency selectivity & speech intelligibility in noise," *Scand Audiol Suppl*, 15, pp. 113 – 122, 1982.

[13] A. Murase, F. Nakajima, S. Sakamoto, Y. Suzuki, and T. Kawase , "Effect of sound localization with dichotic-listening hearing aids," *Proc 18th Int Congress on Acoustics* , Kyoto, Japan, II, pp. 1519 – 1522, 2004.

[14] M. Nilsson, S.D. Soli, and J. A. Sullivan, " Development of the hearing in noise test for the measurement of speech reception thresholds in quiet and in noise," *The Journal of the Acoustical Society of America*, vol. 95(2), pp. 1085-1099, 1994.

[15] M. Nitschmann, J.L. Verhey, and B. Kollmeier B, "Monaural and binaural frequency selectivity in hearing-impaired subjects," *Int J Audiol*., 49(5), pp.357-367, 2010 .

[16] R.D. Patterson, "Auditory filter shapes derived with noise stimuli," *J Acoust Soc Am*, 59(3), pp. 640–654, 1976.

[17] C. V. Pavlovic, ''Derivation of primary parameters and procedures for use in speech intelligibility predictions,'' *J Acoust Soc Am*, 82, pp. 413–422, 1982.

[18] R.W. Peters, and B.C.J. Moore, "Auditory filter shapes at low center frequencies in young and elderly hearing-impaired subjects," *J Acoust Soc Am*, 91(1), pp. 256–266, 1992.

[19] G. E. Peterson, and H. L. Barney, "Control methods in the study of vowels," *J Acoust Soc Am*, 24, pp. 175–184, 1952.

[20] S. Santurette and T. Dau, "Binaural pitch perception in normal-hearing and hearing-impaired listeners," *Hear Res*, 223, pp. 29–47, 2007.

[21] I. W. Selesnick, "Resonance-based signal decomposition: A new sparsity-enabled signal analysis method," *Signal Processing*, vol. 91(12), pp. 2793-2809, December 2011

[22] I. W. Selesnick, "Wavelet transform with tunable Q-factor," *IEEE Transactions on Signal Processing*, vol. 59(8), pp. 3560-3575, August 2011

[23] V. Summers and M.R. Leek, "Intraspeech spread of masking in normal hearing and hearing-impaired listeners," *J Acoust Soc Am*, 101, pp. 2866 – 2876, 1997.

[24] C.T. Tan, B. Guo, and I. Selesnick, "Resonance-based decomposition for the manupulation of acoustic cues in speech: An assessment of perceived quality," *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, New York, USA, pp.333-336, 2011.

[25] R.S. Tyler, J.P. Preece, and M.W. Lowder, The Iowa audiovisual speech perception laser video disc. *Laser Videodisc and Laboratory Report*, University of Iowa at Iowa City, Department of Otolaryngology – Head and Neck Surgery, 1987.

[26] D.J. Van Tasell, "Perception of second-formant transitions by hearing-impaired persons," *Ear Hear*, 1, pp. 130–136, 1980.

[27] Wagener K.C., *Factors influencing sentence intelligibility in noise*, PhD thesis, Carl-von-Ossietzky-Universität, Oldenburg, 2003.