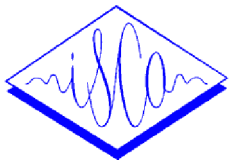

ITRW on Statistical and
Perceptual Audio Processing
Brisbane, Australia
September 21, 2008

Preliminary Intelligibility Tests of a Monaural Speech Segregation System

Ke Hu[1], *Pierre Divenyi*[2], *Dan Ellis*[3], *Zhaozhang Jin*[1], *Barbara G. Shinn-Cunningham*[4], *DeLiang Wang*[1]

[1]Department of Computer Science & Engineering and Center for Cognitive Science
The Ohio State University, Columbus, OH 43210
[2]Speech and Hearing Research, VA Medical Center, Martinez, CA 94553
[3]Department of Electrical Engineering, Columbia University, New York, NY 10027
[4]Departments of Cognitive & Neural Systems and Biomedical Engineering,
Boston University, Boston, MA 02215

huk@cse.ohio-state.edu, pdivenyi@ebire.org, dpwe@ee.columbia.edu,
jinzh@cse.ohio-state.edu, shinn@cns.bu.edu, dwang@cse.ohio-state.edu

Abstract

Human listeners are able to understand speech in the presence of a noisy background. How to simulate this perceptual ability remains a great challenge. This paper describes a preliminary evaluation of intelligibility of the output of a monaural speech segregation system. The system performs speech segregation in two stages. The first stage segregates voiced speech using supervised learning of harmonic features, and the second stage segregates unvoiced speech by subtracting noise energy that is estimated from voiced intervals and onset/offset based segmentation. Objective evaluation in terms of the match to ideal binary time-frequency masks shows substantial improvements. Tests with human subjects indicate that the system improves intelligibility for young listeners when the input SNR is very low, but does not aid elderly listeners. This preliminary evaluation identifies aspects of the system that should be improved in order to produce consistent improvement in intelligibility in noisy environments.

Index Terms: speech segregation, computational auditory scene analysis, ideal binary mask, supervised learning, onset/offset analysis, segmentation

1. Introduction

In real-world listening environments, speech reaching our ears is often corrupted by acoustic interference. The human auditory system segregates a target signal (e.g. speech) from an acoustic mixture using various cues, including fundamental frequency ($F0$), common onset and offset, and amplitude modulation, in the perceptual process called auditory scene analysis (ASA) [1]. Computation auditory scene analysis (CASA) aims to achieve sound organization based on perceptual principles [2]. The $F0$ or pitch cue is widely used in monaural CASA systems; however, systems that employ only this cue are limited to voiced speech segregation. On the other hand, onsets and offsets (corresponding to sudden increases and decreases of signal energy) can be used to segment both voiced and unvoiced speech [3].

Motivated by perceptual and computational considerations, ideal binary mask (IBM) has been suggested as a benmark goal for CASA evaluation [4]. The IBM assigns values of zero and one in the time-frequency (T-F) domain by comparing the local signal-to-noise ratio (SNR) within each T-F unit against a predefined threshold using the source signals that are known a priori. Previous listening tests have shown that speech segregation by IBM leads to dramatic intelligibility improvements [5, 6, 7]. Brungart *et al.* [5] tested IBM processed multitalker mixtures and found that ideal binary masking leads to speech intelligibility improvements on the order of 22-25 dB for normal-hearing listeners. Anzalone *et al.* [6] tested a somewhat different version of IBM, constructed by comparing target speech energy with a fixed threshold, and reported substantial perceptual improvements measured as the reduction in the speech reception threshold (SRT) for both normal-hearing and hearing-impaired listeners. A recent study by Li and Loizou [7] extended the findings of Brungart *et al.* to different types of speech and interference. They also tested the effects of deviation from IBM, and reported that two types of mask errors - misses and false alarms - yield different effects on speech intelligibility.

An IBM is constructed by comparing the sound sources before mixing them together, but in realistic settings, one has to estimate IBM from mixtures directly. In this paper, we describe a monaural speech segregation system that directly estimates the IBM, and preliminary intelligibility tests of this system. The proposed CASA system separates speech from background noise in two separate stages. The first stage is designed to segregate voiced speech, and the second stage segregates unvoiced speech. In the first stage, our system computes a pitch-based grouping cue from a set of harmonic features within each T-F unit, where pitches extracted from premixed speech are used. The transformation from harmonic features to the grouping cue is performed by a multilayer perceptron (MLP) that is trained for each channel of a gammatone filterbank. Voiced segregation lays the foundation for noise estimation, which is then used in the second stage to separate unvoiced speech by spectral subtraction [8]. The second stage of our system also employs an onset/offset analysis in order to further remove residual noise that fails to be eliminated by spectral subtraction. An objective evaluation shows that our system produces good estimates of IBM. Preliminary intelligibility tests, conducted on both young and elderly listeners, indicate that the algorithm works better for young listeners than elderly listeners. In the case of young listeners, the algorithm appears to improve intelligibility most in low SNR condition.

The rest of the paper is organized as follows. The next section gives a detailed description of our system. Objective and subjective evaluations are presented in Section 3 and Section 4, respectively. Concluding remarks are given in Section 5.

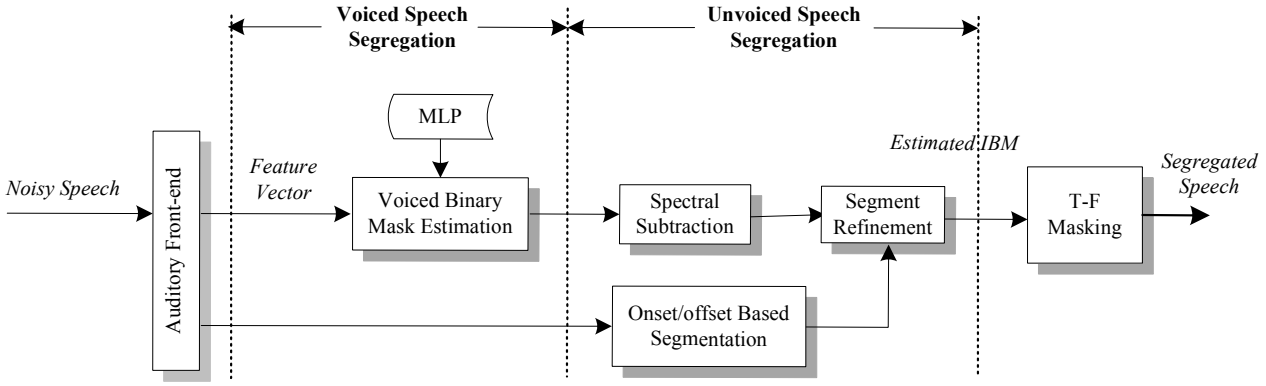

Figure 1: Schematic diagram of the proposed two-stage CASA system. In the voiced-segregation stage, the system uses a set of harmonic features to classify each T-F unit into target or interference. In the unvoiced-segregation stage, the system performs spectral subtraction on noise estimate from the first stage. In addition, onset/offset based segmentation is employed for segment refinement. The resulting time-frequency mask is then used to segregate the target speech.

2. System Description

Our two-stage segregation algorithm is illustrated in Figure 1. The input to the system is a mixture of target speech and background noise. To extract acoustic features, the input mixture is first analyzed using an auditory front-end, which models cochlear filtering and neural transduction. Cochlear filtering is performed using a standard filterbank [9] with 128 frequency channels whose center frequencies are quasi-logarithmically spaced from 50 Hz to 8 kHz. Each channel output is processed by the Meddis model of hair cell to simulate auditory nerve transduction [10]. The output from the Meddis model is divided into 20-ms-long time frames with 10 ms overlap between neighboring frames. The resulting representation is called a cochleagram, and is a first order model of the input signal received by the human auditory system. See [2] for details of cochleagram analysis and synthesis. The envelope of each cochleagram channel is extracted using a bandpass filter (passband from 50 Hz to 550 Hz). The two main stages of the CASA system are described in the following two subsections.

2.1. Voiced speech segregation

In the voiced-speech-segregation stage, an MLP is trained for each frequency channel using a set of noisy utterances. The trained MLP then decides whether the local SNR within a T-F unit exceeds a particular threshold [11]. Specifically, within each T-F unit we extract a vector of harmonic features as the input to the MLP.

Following [11, 12], a 6-dimensional feature vector is extracted to represent the harmonic structure of the T-F unit of frequency channel c and time frame m in voiced intervals (sets of consecutive voiced frames):

$$\mathbf{x}_{c,m} = \begin{pmatrix} A(c,m,\tau_m) \\ [\bar{f}(c,m)\cdot\tau_m] \\ \left|\bar{f}(c,m)\cdot\tau_m - [\bar{f}(c,m)\cdot\tau_m]\right| \\ A_E(c,m,\tau_m) \\ [\bar{f}_E(c,m)\cdot\tau_m] \\ \left|\bar{f}_E(c,m)\cdot\tau_m - [\bar{f}_E(c,m)\cdot\tau_m]\right| \end{pmatrix}, \quad (1)$$

where τ_m is the pitch period for frame m. $A(c,m,\tau_m)$ is the autocorrelation of the front-end response with time lag τ_m and $A_E(c,m,\tau_m)$ is the envelope autocorrelation of the response. $\bar{f}(c,m)$ denotes the estimated average instantaneous frequency of the unit response, and $\bar{f}_E(c,m)$ its envelope. $[\cdot]$ denotes the round-to-integer operation. The first three features are extracted from auditory front-end responses, and the last three features from the envelopes of the front-end responses. Each set of three features captures the periodicity, harmonic number, and deviation from the nearest harmonic, respectively for one T-F unit.

Given the feature vector, we train an MLP for each channel to directly maximize the SNR of segregated speech using the following objective function

$$J_c = \Sigma_m (d_c(m) - y_c(m))^2 \cdot E_c(m) / \Sigma_m E_c(m), \quad (2)$$

where $E_c(m)$ denotes the energy at frame m and channel c, and $d_c(m)$ and $y_c(m)$ the desired (binary) and actual outputs. It is worth mentioning that J_c is a generalized form of mean square error, with each error term weighted by the normalized energy within the corresponding T-F unit. In implementation, we generalize the Levenberg-Marquardt backpropagation algorithm [13] to minimize J_c.

All the MLP's have the same architecture, with one hidden layer of 20 nodes. During MLP training, IBM provides the desired output. Our informal listening indicates that, when the mixture SNR is below zero, the IBM created using a local SNR criteria (LC) [5] that is equal to the mixture SNR seems to produce better intelligibility than fixing LC to 0 dB. Thus, in IBM construction we choose LC to be 0 dB when the mixture SNR is greater than or equal to 0 dB and equal to the mixture SNR when the latter is less than 0 dB. Each trained MLP is then used to label only the T-F units of its corresponding channel in voiced intervals. A T-F unit is labeled as target speech if the posterior probability that the unit contains relatively strong target energy is greater than the posterior probability that the unit contains relatively strong interference energy.

2.2. Unvoiced speech segregation

Because the feature vector in (1) encodes harmonic structure, it cannot segregate unvoiced speech. The second stage of our system deals with the unvoiced speech segregation problem from a different perspective. Specifically, we estimate the interference within voiced intervals by capitalizing on the results of the first

Table 1: SNR gain (in dB) of the proposed system across four SNR conditions

Mixture SNR	Voiced speech SNR gain	Unvoiced speech SNR gain	Overall SNR gain
-2	10	19.5	12.2
-4	10.4	20.7	12.6
-6	10.9	21.7	13
-8	11.5	22.5	13.6

stage [12], and then apply spectral subtraction to remove interference during unvoiced intervals, hence segregating unvoiced speech. Specifically, the system first estimates noise energy for each channel from the previous voiced interval by averaging mixture energy in the T-F units labeled as interference in the first stage; if such a voiced interval does not exist (which can happen at the beginning of an utterance) the succeeding voiced interval is used for noise estimation. Given estimated noise energy, we estimate the local SNR in each T-F unit in the current unvoiced interval. A T-F unit is labeled as target speech if and only if its local SNR is above LC.

We find that just applying spectral subtraction still retains many T-F units dominated by interference. To further remove residual noise after subtraction, we employ multi-scale onset/offset based segmentation [3] to refine the results from spectral subtraction. To conduct onset/offset based segmentation, we first detect onsets and offsets in each frequency channel. We then align detected onsets neighboring in frequency and close in time to form an onset front; detected offsets are aligned to form an offset front. A segment is extracted from a matching pair of onset and offset fronts. The final set of segments is produced by integrating over several analysis scales. This set of segments is then compared with segments produced by simply merging neighboring T-F units labeled by spectral subtraction. If a subtraction-based segment overlaps with an onset/offset based segment so that at least 90% of the latter energy is contained in the overlapping region, the subtraction-based segment is kept; otherwise, it is removed. This segment refining serves to remove isolated fragments that likely belong to background noise.

3. Objective Evaluation

In this section, we evaluate our system on mixtures of IEEE sentences embedded in matched speech-shaped noise (SSN) [14]. The IEEE sentence corpus contains 720 phonetically-balanced sentences with relatively low word-context predictability. All sentences were recorded by a single female speaker at a 20 kHz sampling frequency. We downsample the signals to 16 kHz. Mixtures of the IEEE sentences and SSN are created at four SNR conditions (-2 dB, -4 dB, -6 dB, and -8 dB) in order to match the test conditions of perceptual tests (described in the next section). In each SNR condition, 100 mixtures are used for MLP training and the remaining 620 are used to evaluate system performance. Feature extraction requires the knowledge of $F0$ at each frame, so we use pitch contours extracted from premixed speech utterances using Praat [15] in order to remove the influence of pitch estimation errors on segregation performance.

Given that the computational objective of our proposed system is to estimate IBM, we adopt the same SNR measure in [16] and use the resynthesized speech from the ideal binary mask as the ground truth

$$SNR = 10\log_{10}(\Sigma_n S_I^2[n]/\Sigma_n(S_I[n] - S_E[n])^2) \quad (3)$$

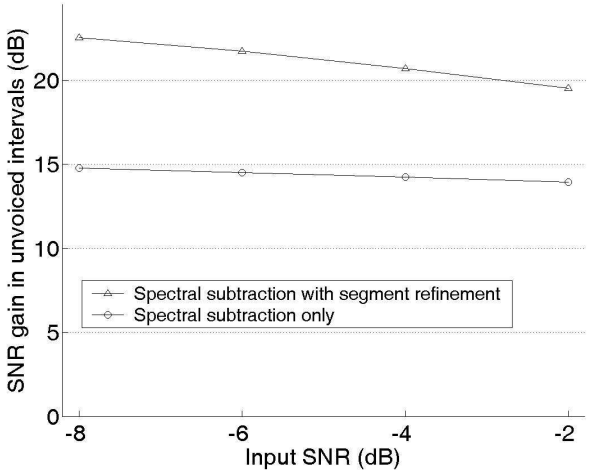

Figure 2: SNR gain comparison for the proposed system in unvoiced intervals with and without segment refinement.

where $S_I[n]$ and $S_E[n]$ are signals resynthesized from IBM and the mask estimated by our segregation system, respectively.

Table 1 summarizes the performance of our system in terms of SNR gain at different input SNR levels. We report the results of voiced speech segregation only (the first stage), unvoiced speech segregation only (the second stage) and then the entire system in the table. Our system obtains an overall SNR gain of 13.6 dB when the input SNR is -8 dB. The gain drops slightly to 12.2 dB when the input SNR is increased to -2 dB. Although the voiced segregation performance is measured within voiced intervals only, it is a little lower than the overall segregation performance, presumably because the input SNR's in voiced intervals are higher than in unvoiced intervals. Thus, SNR gain tends to be smaller as the input SNR increases. On the other hand, the SNR gains in unvoiced intervals are significantly higher than the overall SNR improvements.

To isolate the effects of segment refinement using onset/offset based segmentation, Figure 2 shows the SNR gain in unvoiced intervals with and without segment refinement following spectral subtraction. As shown in the figure, onset/offset segmentation is responsible for a substantial amount of SNR improvement.

Li and Loizou [7] observed that intelligibility score drops gradually as the percentage of wrongly labeled T-F units increases. Motivated by their results, we further evaluate our system performance in terms of error percentages in unit labeling. First, the overall percentage of mask error is calculated as the average error rate per frame, counting flips from 0's to 1's and from 1's to 0's, relative to the IBM. These error rates are given in Table 2. Second, the two different types of possible error – misses and false alarms – have been shown to have different

Table 2: Error rates (in %) produced by the proposed system in estimating IBM

Input SNR	Miss error	False alarm error	Overall error
-2 dB	25.61	4.26	7.82
-4 dB	29.72	4.19	8.87
-6 dB	34.98	4.11	10.15
-8 dB	39.30	4.07	11.35

impacts on speech intelligibility [7]. Therefore, we examine the two types of error separately. Specifically, the miss error is calculated as the per-frame average percentage of target units (1's) wrongly labeled as interference (0's), the false alarm error is defined as the per-frame average percentage of interference units wrongly labeled as target. Rates of these two types of error are shown in Table 2. Notice that the overall error in [7] is created according to a pre-defined percentage, and thus has zero variance. We have further analyzed the error histogram and found that more than 95% of frames have mask error rates lower than 20%. In comparison with the overall rates of representative speech enhancement systems [7], our algorithm achieves considerably lower error rates across all SNR conditions. Also, our system makes more miss errors than false alarm errors. It has been shown that miss errors are more benign than false alarm errors for speech intelligibility [7].

According to the relationship between overall error and intelligibility observed in [7] for IEEE sentences mixed with SSN, one would predict that the proposed system should lead to significantly higher intelligibility over no processing.

4. Preliminary Intelligibility Tests

4.1. Methods

4.1.1. Stimuli

Test signals were created as described in Section 3. For each test sentence and each mixture SNR level, two test stimuli were generated: a segregated mixture that is the output of our system and an unsegregated mixture. To account for filtering effects and any possible distortion introduced during cochleagram analysis and synthesis, an input mixture in the unsegregated case is processed through an all-1 mask.

4.1.2. Subjects

Subjects were recruited at the Speech and Hearing Research Lab of the VA Medical Center in Martinez, California. One group of five young listeners (between 21 and 33 years of age) and another group of four elderly listeners (between 76 and 83 years of age) participated in the preliminary tests. All young subjects had normal hearing (pure tone thresholds under 5 dB HL [hearing level] between 0.5 and 4 kHz). The elderly subjects had mild-to-moderate hearing loss (0.5 to 4-kHz pure tone average thresholds between 12 and 39 dB HL in the better ear and between 12 and 41 dB HL in the worse ear); none of the subjects had ever used a hearing aid. Since the presentation was diotic, the effective HL of each subject should be regarded as that of the better ear.

4.1.3. Procedure

Subjects were seated in a sound-attenuated booth and listened to the material presented diotically through earphones (Sennheiser HD-580). Segregated utterances were presented through a GINA-24 subsystem, which are peripherals of experimental PC's, at a comfortable listening level (both the average level of unsegregated mixtures and the average level of vowel segments of segregated mixtures were presented at a fixed level of 72 dB SPL at the earphone). The subjects' task was either to type the sentence heard on a computer keyboard or, as the majority of the elderly subjects chose to do, to repeat the sentence to the experimenter who typed the response for them. Obvious typographic errors were ignored. The sentence material was presented in lists of ten. Between lists the subjects were allowed to rest as they wished.

4.2. Results and discussion

Because a large number of studies conclude that the given performance of speech understanding in noise requires a higher SNR for elderly than for young listeners, we tested the two groups at different SNR's. For the young subjects, both unsegregated and segregated mixtures were presented at -8, -6, and -4 dB SNR; the elderly subjects were presented the material at -4 and -2 dB SNR. Since the IEEE material consists of grammatically correct but not easily predictable sentences, each of which contains five keywords, intelligibility was measured as the percentage of correct keywords.

Intelligibility results of the young listeners are given in Table 3. In the table we list the intelligibility of the unsegregated and segregated conditions separately. As observed from Table 3, all young subjects scored higher in recognizing segregated speech than the unsegregated one at the input SNR of -8 dB. On the other hand, some young listeners obtained improvement at -6 dB and -4 dB while others performed worse in the segregated conditions, leading to almost equal average intelligibility. These results, although preliminary and with only a small number of subjects, indicate the effectiveness of our algorithm in very low SNR situations.

Intelligibility test results of the four elderly listeners are shown in Table 4. We observe that, at input SNR levels of -2 dB and -4 dB, there is no improvement in the intelligibility of segregated mixtures; indeed the average score even dropped as a result of segregation. When asked of the difficulty in understanding segregated speech, all elderly listeners reported having been disturbed by the extraneous chirps and whistles inherent in binary-masked mixtures. These reports are actually consistent with previous observations showing that the elderly are more easily distracted from listening to speech targets [17], possibly as the result of age-related suppression of inhibitory processes [18]. The chirps and whistles that they heard can be characterized as musical noise, which has been identified as one of main reasons why speech enhancement algorithms fail to improve speech intelligibility. Such noise seems to be more disturbing for the elderly than the young.

By examining IBM estimation errors in Table 2, one could arrive at an estimate of intelligibility score for normal-hearing listeners according to the relationship between intelligibility and mask error reported in [7]. As 95% of the frames in segregated mixtures have overall mask error rates smaller than 20%, our system would be expected to achieve a percent-correct score between 70% and 90% when the mixture SNR is around -5 dB. There are several possible reasons for the gap between the expected intelligibility and the observed results in Table 3 and Table 4. First, our IBM construction uses different LC values rather than the 0 dB uniformly used in [7]. Second, we use cochleagrams as the signal representation whereas they used

Table 3: Percent correct keyword intelligibility as a function of input SNR for five young subjects. The mean and standard deviation are given in the bottom row

Subject		Input SNR		
		-4 dB	-6 dB	-8 dB
Y1	Unsegregated	35.6	16.0	12.4
	Segregated	42.8	36.0	16.0
Y2	Unsegregated	42.4	28.0	12.8
	Segregated	52.0	34.4	20.0
Y3	Unsegregated	41.2	32.0	9.6
	Segregated	44.0	30.8	17.2
Y4	Unsegregated	53.2	50.0	26.0
	Segregated	46.8	33.6	26.4
Y5	Unsegregated	77.6	59.6	24.4
	Segregated	62.8	52.0	34.3
Mean (± Std)	Unsegregated	50.0 (±16.7)	37.1 (±17.5)	17.0 (±7.6)
	Segregated	49.7 (±8.1)	37.4 (±8.4)	22.8 (±7.6)

Table 4: Percent correct keyword intelligibility as a function of input SNR for four elderly subjects. The mean and standard deviation are given in the bottom row

Subject	Keyword Accuracy	Input SNR	
		-2 dB	-4 dB
E1	Unsegregated	55.6	35.2
	Segregated	37.2	27.2
E2	Unsegregated	76.8	70.8
	Segregated	53.6	46.8
E3	Unsegregated	64.4	48.0
	Segregated	58.4	46.0
E4	Unsegregated	72.4	62.0
	Segregated	71.2	56.4
Mean (± Std)	Unsegregated	67.3 (±9.3)	54.0 (±15.6)
	Segregated	55.1 (±14.1)	44.1 (±12.2)

spectrograms. Third, our tests were conducted with fewer subjects and different experimental protocols. Further research is clearly needed in order to understand the apparent discrepancy.

5. Concluding Remarks

In this paper, we presented a preliminary intelligibility study of a CASA system for speech segregation. The system first separates voiced speech using MLP trained on harmonic features. Speech segregation in voiced intervals provides the basis for estimating noise energy, which is then used to segregate unvoiced speech. Objective evaluation shows that our system obtains substantial SNR improvements. The intelligibility study further indicates that our algorithm can improve speech intelligibility under very low SNR conditions. For elderly listeners, processing artifacts in binary masking appear to contribute to worse intelligibility of segregated mixtures.

We emphasize the preliminary nature of our intelligibility tests. The small subject pools inevitably led to large performance variability. We do not fully understand why the observed intelligibility seems lower than would be predicted from the observed relationship between intelligibility and mask error by Li and Loizou [7]. It is unclear why elderly listeners did not perform worse than young listeners in unsegregated conditions, as previous studies suggest. Also, our results seem inconsistent with the results of Anzalone *et al.* [6] who found larger SRT improvements for listeners with hearing loss (who are also elderly) than for normal-hearing subjects, although most of their subjects were experienced hearing aid wearers. Future tests are planned for more subjects and more systematic test protocols.

On the side of system development, our use of known $F0$ needs to be replaced by a pitch tracking algorithm for noisy speech. Our method of unvoiced speech segregation implicitly relies on the assumption that noise energy does not change much from a voiced interval to its succeeding unvoiced interval, which works well for approximately stationary intrusions such as SSN, but is clearly violated in multitalker scenarios.

Given the long-standing challenge of improving speech intelligibility in monaural speech separation [19, 20], the fact that our system shows some improvement in low SNR conditions is encouraging. Our preliminary tests also point to ways of potentially making segregated mixtures more intelligible. For example, methods of attenuating musical noise caused by binary

masking [21, 22] may be helpful, particularly for elderly listeners. Also, different choices of LC values for IBM construction could lead to better intelligibility performance.

6. Acknowledgement

The authors wish to express their gratitude to Casey Knifsend and Adam Lammert for supervising the listening tests. This research was supported in part by an NSF Collaborative Grant (IIS-0534707) and the VA Biomedical Laboratory Research and Development Program.

7. References

[1] S. Bregman, *Auditory Scene Analysis*. Cambridge, MA: MIT press, 1990.

[2] D. L. Wang and G. J. Brown, Eds., *Computational auditory scene analysis: Principles, algorithms and applications*. IEEE Press/Wiley-Interscience, 2006.

[3] G. Hu and D. L. Wang, "Auditory segmentation based on onset and offset analysis," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, pp. 396–405, 2007.

[4] D. L. Wang, "On ideal binary mask as the computational goal of auditory scene analysis," *Speech separation by humans and machines*, P. Divenyi, Ed., Norwell, MA: Kluwer Academic 2005, pp. 181-197.

[5] D. S. Brungart, P. S. Chang, B. D. Simpson, and D. L. Wang, "Isolating the energetic component of speech-on-speech masking with ideal time-frequency segregation," *Journal of the Acoustical Society of America*, vol. 120, pp. 4007–4018, 2006.

[6] M. C. Anzalone, L. Calandruccio, K. A. Doherty, and L. H. Carney, "Determination of the potential benefit of time-frequency gain manipulation," *Ear and Hearing*, vol. 27(5), pp. 480–492, 2006.

[7] N. Li and P. C. Loizou, "Factors influencing intelligibility of ideal binary-masked speech: Implications for noise reduction," *Journal of Acoustic Society of American*, vol. 123, pp. 1673–1682, 2008.

[8] S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Acoust., Speech, Signal Process*, vol. 27, pp. 113–120, 1979.

[9] R. D. Patterson, J. Holdsworth, I. Nimmo-Smith, and P. Rice, "SVOS final report, part b: Implementing a gammatone filterbank. rep. 2341," MRC Applied Psychology Unit, Tech. Rep., 1988.

[10] R. Meddis, "Simulation of auditory-neural transduction: Further studies," *Journal of Acoustic Society of American*, vol. 83, pp. 1056–1063, 1988.

[11] Z. Jin and D. L. Wang, "A supervised learning approach to monaural segregation of reverberant speech," in *Proceedings of IEEE-ICASSP*, 2007, pp. IV.921–924.

[12] G. Hu, "Monaural speech organization and segregation," Ph.D. dissertation, Ph.D. dissertation, Biophysics Program, The Ohio State University, 2006.

[13] M. T. Hagan, H. B. Demuth, and M. H. Beale, *Neural Network Design*. Boston, MA: PWS Publishing, 1996.

[14] IEEE, "IEEE recommended practice for speech quality measurements," *IEEE Trans. Audio and Electroacoustics*, vol. 17, pp. 225–246, 1969.

[15] P. Boersma and D. Weenink, "Praat: doing phonetics by computer," 2005. Available at http://www.fon.hum.uva.nl/praat.

[16] G. Hu and D. L. Wang, "Monaural speech segregation based on pitch tracking and amplitude modulation," *IEEE Trans. Neural Networks*, vol. 15, pp. 1135–1150, 2004.

[17] P. A. Tun, G. O'Kane, and A. Wingfield, "Distraction by competing speech in young and older adult listeners," *Psychology and Aging*, vol. 17, pp. 453–467, 2002.

[18] M. Pilotti, T. Beyer, and M. Yasunami, "Top-down processing and the suffix effect in young and older adults," *Memory & Cognition*, vol. 30, pp. 89–96(8), 2002.

[19] H. Levitt, "Noise reduction in hearing aids: A review," *Journal of Rehabilitation Research and Development*, vol. 38, pp. 111–121, 2001.

[20] H. Dillon, *Hearing aids*. New York: Thieme, 2001.

[21] S. Araki, S. Makino, H. Sawada, and R. Mukai, "Reducing musical noise by a fine-shift overlap-and-add method applied to source separation using a time-frequency mask," in *Proceedings of IEEE-ICASSP*, 2005, pp. III.81–84.

[22] C. Breithaupt, T. Gerkmann, and R. Martin, "A noval a priori snr estimation approach based on selective cepstro-temporal smoothing," in *Proceedings of IEEEICASSP*, 2008.