To be Published in: *Proceedings of IEEE International Workshop on Neural Networks for Signal Processing* (NNSP'03), 2003

SEGREGATION OF STOP CONSONANTS FROM ACOUSTIC INTERFERENCE

Guoning Hu

Biophysics Program
The Ohio State University
Columbus, OH43210, USA
hu.117@osu.edu

DeLiang Wang

Department of Computer and Information Science & Center of Cognitive Science
The Ohio State University
Columbus, OH43210, USA
dwang@cis.ohio-state.edu

ABSTRACT

Speech segregation from acoustic interference is a very challenging task. Previous systems have dealt with voiced speech with success, but they cannot handle unvoiced speech. We study the segregation of stop consonants, which contain significant unvoiced signals. We propose a novel method that employs onset as a major cue to segregate stop consonants. Our system first detects stops through onset detection and Bayesian classification of acoustic-phonetic features, and then performs grouping based on onset coincidence. The system has been tested and performs well on utterances mixed with various types of interference.

1. INTRODUCTION

The segregation of speech from acoustic interference is required in many applications, such as speech recognition and hearing aids design. Currently, no method performs this task well in real environments. Listeners extract speech signals through a process called auditory scene analysis (ASA), in which the time-frequency regions dominated by the target speech are identified and grouped into a stream [2]. Previous speech separation efforts based on ASA principles utilize harmonicity as the major ASA cue, hence are limited to voiced speech [3] [4]. To deal with unvoiced speech, other ASA cues must be explored.

In this paper, we address the problem of separating stop consonants from interference. Stop consonants contain /t/, /d/, /p/, /b/, /k/, and /g/, which occur frequently in natural speech. A stop usually starts with a closure [10], which corresponds to the stop of airflow in the vocal tract, and a burst, which corresponds to a sudden release of air. As an example, the waveform of a stop /g/ and its spectrogram are shown in Fig. 1(a) and 1(b). Consistent with ASA principles [2], our objective is to identify the time-frequency regions where stop sounds are dominant, and grouping them into a target stream. Because the closure

of a stop consonant contains little acoustic information, we will focus on separating stop bursts.

Since the acoustic realization of a stop burst is mainly unvoiced, it cannot be separated based on harmonicity. Nevertheless, at its onset, a significant intensity increase happens across a wide frequency range (see Fig. 1(b)). Therefore, we can identify stop bursts by detecting their onsets, and then group them based on onset coincidence. Note that onset is an important ASA cue [2]. To detect the onsets of stop bursts, an acoustic mixture is first decomposed through an auditory filterbank. Then an onset detector identifies stop candidates by detecting local onsets in each filter channel and integrating information across all the channels. Finally, to distinguish true stop bursts from burst-like interference, these stop candidates are further classified through a Bayesian decision rule on auditory-acoustic features. Here three features are employed: burst duration, auditory spectrum, and relative intensity. Prior probabilities for these features are obtained from a training dataset, which contains all utterances from the training set of TIMIT database for stops, and 18 natural sounds for interference.

This paper is organized as follows. Sect. II and Sect. III describe details of stop detection and classification. Sect. IV presents the results of classification and grouping. A brief discussion is given in Sect. V.

2. ONSET DETECTION

The input signal is sampled at 16 kHz and normalized into 80 dB sound pressure level. It is first analyzed by a model of auditory periphery, which includes cochlear filtering and auditory nerve transduction. We use an auditory filterbank with 150 gammatone filters [7] centered from 80 Hz to 7 kHz to model cochlear filtering,

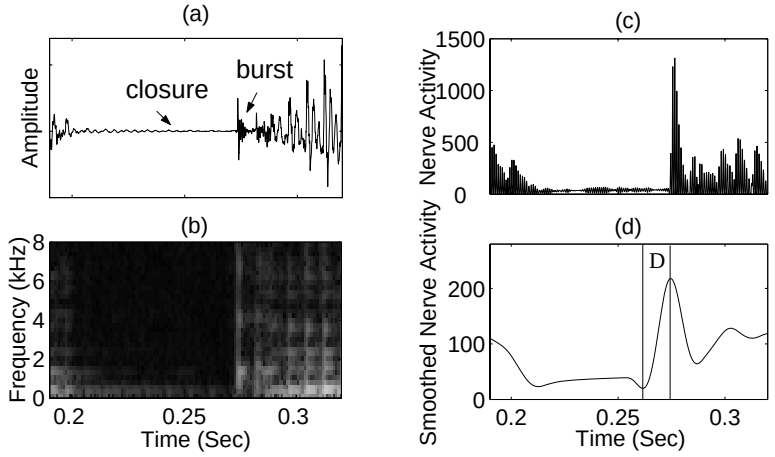

Figure 1. Waveform (a) and spectrogram (b) of /g/ in word “good”; corresponding auditory nerve activity (c) and smoothed nerve activity (d) in a channel centered at 1 kHz.

and the Meddis model for neural transduction [5]. The output, in the form of auditory nerve activity, is decomposed into 20 ms frames with 10 ms frame shift.

We first detect onsets in each filter channel. The Meddis model exhibits saturation and fast adaptation that can be conveniently used for onset detection. An onset of an acoustic event in a frequency band corresponds to a large intensity increase in the response of the corresponding gammatone filter. Following the increase, the auditory nerve activity increases rapidly to a significant level and then decreases to a steady state. As an example, Fig. 1(c) shows the nerve activity from a channel centered around 1 kHz for the input in Fig. 1(a). Based on the above analysis, we propose to detect onsets as follows. First, to remove some small intensity fluctuations, each auditory nerve response is smoothed with a lowpass filter, and its first derivative is computed. Since a large increase corresponds to a large derivative, we identify onsets by marking peaks that exceed a certain threshold (to eliminate spurious peaks due to small intensity fluctuations that do not correspond to onsets). Our onset detector is similar to the standard Canny edge detector in visual processing, in which Gaussian smoothing performs lowpass filtering [8].

Here, we use a lowpass filter with transition band from 30 Hz to 80 Hz. Its passband ripple and stopband ripple are 0.1 and 0.02, respectively. Fig. 1(d) illustrates the smoothed nerve activity corresponding to Fig. 1(c). Let $a(c,t)$ be the smoothed auditory nerve response in channel c at time t. Its derivative is approximated by $a(c,t) - a(c,t-\tau)$. τ is a constant that corresponds to the average increase period of the nerve activity for stop bursts, the duration from a local maximum of a to the preceding local minimum (D in Fig. 1(d)). From the training set, we obtain τ = 14.375 ms. Since the derivative corresponding to an onset is generally greater than the difference between the average steady-state nerve activity and the spontaneous nerve activity (see [5] for more details), peaks above this difference are marked as channel onsets.

For stops in the training set, except for a few weak stops, they trigger onsets in 10 or more adjacent channels simultaneously. Therefore, when 10 or more adjacent channels have onsets at a particular time, the detector will identify a stop candidate there.

3. STOP CLASSIFICATION

Since detected onset candidates may correspond to burst-like sounds from interference, they are classified based on auditory-acoustic features. Let H_0 denote a hypothesis that a candidate is from interference, and H_j a true stop. Here j = 1, 2, …, 6, corresponding to /t/, /d/, /p/, /b/, /k/, and /g/, respectively. Let $\mathbf{X}$ be the feature vector for a stop candidate, and $p(H_j|\mathbf{X})$ the posterior probability of H_j given $\mathbf{X}$, for j = 0, 1, …, 6. According to the Bayesian decision rule, the candidate is classified as a stop if $p(H_0|\mathbf{X})$ is not the maximum among them. Since our objective here is to distinguish true stops from bursts from interference, we do not treat it as an error if a stop is classified into another type of stop. As a result, let H_S denote a hypothesis that a candidate is from a stop. To achieve the minimum

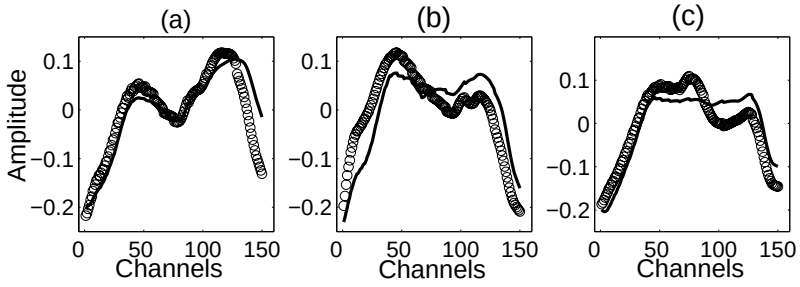

Figure 2. The average auditory spectra of stops: (a) Circle: /t/; line: /d/ (b) Circle: /p/; line: /b/ (c) Circle: /k/; line: /g/.

error rate of classification, a candidate is classified as a stop if and only if $p(H_0|\mathbf{X}) < p(H_S|\mathbf{X})$. Applying Bayesian formula, we have:

$$\frac{p(H_0 \mid \mathbf{X})}{p(H_S \mid \mathbf{X})} = \frac{p(\mathbf{X} \mid H_0)p(H_0)}{p(\mathbf{X} \mid H_S)p(H_S)} \tag{1}$$

The key to construct a good classifier is to choose appropriate features. Previous research suggested that the following features characterize stops: formant transitions, burst spectrum, burst amplitude, durations, and voicing of the closure (see [1] for example). Since our main goal is to separate onsets from interference, we shall choose the distinctive features that are robust to acoustic interference.

We use the burst duration as the first feature, which is obtained as follows. First, we define the auditory spectrum at time t, $\mathbf{S}(t)$, as:

$$\mathbf{S}(t) = (a(1,t), a(2,t), \cdots, a(150,t)) . \tag{2}$$

We call $\mathbf{S}(t)$ auditory spectrum since it is obtained from the output of the auditory filterbank. For a stop candidate m, let t_m be the time where a stop candidate is identified, and $T(t_m)$ the time interval centered around t_m so that for any $t \in T(t_m)$, the cross-correlation between $\mathbf{S}(t)$ and $\mathbf{S}(t_m)$ is higher than 0.6. That is,

$$\hat{\mathbf{S}}(t) \bullet \hat{\mathbf{S}}(t_m) > 0.6 , \quad t \in T(t_m) . \tag{3}$$

Here, $\hat{\mathbf{S}}(t)$ is the normalized auditory spectrum, which has zero mean and unity variance. The burst duration, d_m, is the length of $T(t_m)$.

Each stop phoneme has a particular articulatory gesture, which gives it unique spectral characteristics [10]. For each stop phoneme, the average auditory spectrum within the training set, as shown in Fig. 2, is obtained to capture its spectral characteristics. Note that phonemes with the same place of articulation

[10] have similar average auditory spectra. For a stop candidate *m*, the cross-correlations between its average auditory spectrum in $T(t_m)$ and these templates quantify their similarities. The six cross-correlations are denoted by $\mathbf{c}_m = (c_{m,1}, c_{m,2}, \ldots, c_{m,6})$, corresponding to /t/, /d/, /p/, /b/, /k/, and /g/, respectively.

The intensity of a stop burst is related to the intensity of neighboring voiced speech [10], while the intensity of interference is generally independent from speech. Let $I(m)$ be the average intensity of a candidate *m*, and I_V the average intensity of the input signal in the nearest voiced portion. The relative intensity of it, denoted by r_m, is defined as:

$$r_m = 10\log_{10}[I(m)/I_V]. \tag{4}$$

To compute $I(m)$, the intensity of the output from every channel (gammatone filter) in $T(m)$ is calculated. The channel with the highest intensity is selected, and $I(m)$ is the average intensity of the 10 adjacent channels centered at the selected channel. I_V is computed in a similar way.

We use these three features for classification. More specifically, for a stop candidate *m*, we first use maximum likelihood to identify the stop phoneme *j** that is most similar to candidate *m* according to (d_m, $c_{m,j}$, r_m). That is,

$$j^* = \arg\max_j p(d_m, c_{m,j}, r_m \mid H_j), \quad j = 1, 2, \cdots, 6. \tag{5}$$

Note that the above decision does not determine whether the candidate comes from interference. Then, $\mathbf{X}_m = (d_m, c_{m,j^*}, r_m)$ is used as the feature vector for classification in (1). For simplicity, we approximate $p(\mathbf{X}_m|H_S)$ with $p(\mathbf{X}_m|H_{j^*})$. Therefore we have:

$$\frac{p(\mathbf{X}_m \mid H_0)}{p(\mathbf{X}_m \mid H_S)} \approx \frac{p(d_m, c_{m,j^*}, r_m \mid H_0)}{p(d_m, c_{m,j^*}, r_m \mid H_{j^*})}. \tag{6}$$

To estimate the likelihood, let N_j be the set containing all the stop phoneme *j* from the training set. We find that the distribution of a certain feature *x* within N_j cannot always be approximated well using a model-based approach. Therefore, we estimate $p(x|H_j)$ through the following Kernel estimation [9]:

$$p(x \mid H_j) = \frac{1}{n}\sum_{i \in N_j} \frac{1}{h} K\left(\frac{x - x_i}{h}\right). \tag{7}$$

Here, *n* is the size of the set N_j and *K* is a Gaussion Kernel. *h* is the smoothing parameter obtained as follows:

$$h = \left(\frac{4}{3n}\right)^{1/5} \sigma_x, \tag{8}$$

where σ_x is the variance for feature *x* of the samples within set N_j. For more details, see [9]. Through (7), we estimate $p(d|H_j)$, $p(r|H_j)$, and $p(c_j|H_j)$, for j = 1, 2,

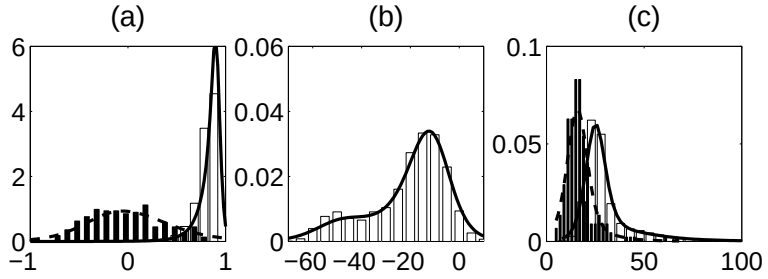

Figure 3. (a) White bar: the histogram of c_1 for stop /t/; black bar: the histogram of c_1 for interference; solid line: estimated $p(c_1|H_1)$, solid line: estimated $p(c_1|H_0)$. (b) White bar: the histogram of r for stop /t/; solid line: estimated $p(r|H_1)$. (c) White bar: the histogram of d for stop /t/; black bar: the histogram of d for interference; solid line: estimated $p(d|H_1)$, dash line: estimated $p(d|H_0)$.

…, 6. Similarly, $p(d|H_0)$ and $p(c_j|H_0)$ are estimated from candidates from the interference in the training set.

As an example, the histograms and estimated likelihood of d, c, and r for /t/, and those of d and c for interference in the training set are shown in Fig. 3. In addition, since r is the ratio between the intensity of a candidate and voiced speech, we cannot estimate $p(r|H_0)$ from interference in the training set. Therefore, we simply use a uniform distribution from –70 dB to 10 dB as $p(r|H_0)$ with the following considerations. First, the onset detector generally will not be sensitive to signals that are more than 70 dB below voiced speech. Second, stops are seldom more than 10 dB above voiced speech.

To determine their independence, the mutual information among these features is computed. The obtained mutual information is very small for each hypothesis. As a result, we treat d, c, and r as independent given each hypothesis, which will greatly simplify the problem. Finally, for a candidate m, we have

$$\frac{p(H_0 \mid \mathbf{X}_m)}{p(H_S \mid \mathbf{X}_m)} = \frac{p(d_m \mid H_0)\, p(c_{m,j^*} \mid H_0)\, p(r_m \mid H_0)\, p(H_0)}{p(d_m \mid H_{j^*})\, p(c_{m,j^*} \mid H_{j^*})\, p(r_m \mid H_{j^*})\, p(H_S)}. \tag{9}$$

The ratio of $p(H_0)$ and $p(H_S)$ varies considerably under different circumstances. For simplicity, we estimate $p(H_S)$ as the average number of detected stops per second of unvoiced speech, and $p(H_0)$ the average number of detected candidates per second over different interference. From the training set, approximately we have $p(H_0)\,/p(H_S) = 1$.

The transition between a stop burst and the following voiced phoneme provides useful information and could be used as another feature. However, the formant transition from a stop to the following voiced phoneme is very difficult to obtain. In addition, it is closely related to the burst spectrum. The voicing of the closure is not robust when interference is strong. Therefore, they are not employed.

4. RESULTS

Detected stops are segregated through grouping signals starting simultaneously with them. More specifically, for a detected stop *m*, if any channel contains an onset, a speech-dominant area in this channel is identified. This area is from the local minimum of smoothed nerve activity preceding $T(t_m)$ to the local minimum of smoothed nerve activity following $T(t_m)$. It contains the major part of the burst since a local minimum of smoothed nerve activity generally corresponds to the onset or the offset of an acoustic event. Time-frequency (T-F) units overlapping with this area are marked as speech dominant. A T-F unit corresponds to input signal in a certain channel and at a certain frame. A binary mask is constructed by assigning 1 to a marked T-F unit and 0 otherwise. The segregated stop signal is resynthesized through the binary mask, which retains the acoustic energy from the mixture corresponding to 1's and rejects that corresponding to 0's (see [3] for more details).

We tested the above method with 10 utterances, which are randomly chosen from the test part of TIMIT database, mixed with 10 intrusions: white noise, pink noise, airplane noise, car noise, factory noise, noise burst, clicks, bar noise, a firework show, and rain. Neither the utterances nor the intrusions are included in the training set. To evaluate the performance, let E_M be the percentage of missing stops, which is the percentage of undetected stops among all the stops. Let E_F be the percentage of false detection, which is the percentage of bursts from interference among all the detected stops. E_M and E_F at different overall SNR levels are shown in Table 1. E_M increases significantly as SNR decreases since stops are more seriously corrupted as interference becomes stronger, while E_F increases moderately. Note that the Bayesian classifier is designed to distinguish

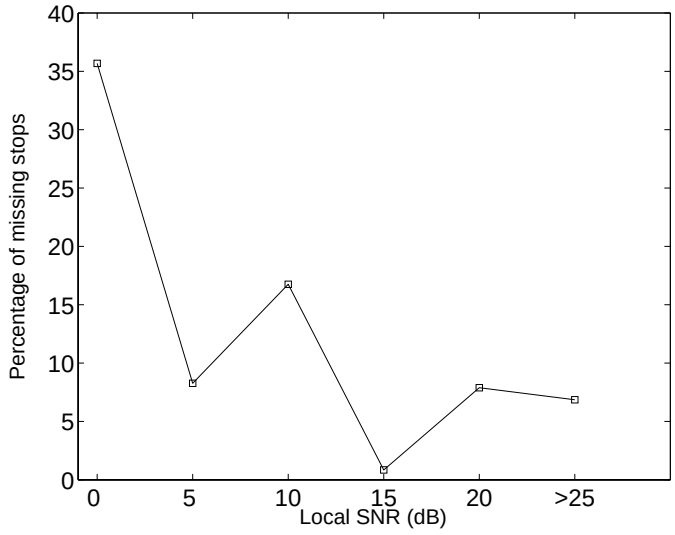

Figure 4. The percentage of missing stops with respect to local SNR

stops from interference. Therefore, we do not count detected bursts that are actually onsets of other phonemes in target speech when calculating E_F. This type of error is not harmful for speech separation since it in essence includes speech signals other than stop consonants, while the goal of speech separation is to remove interference.

To gain more insight into the performance related to the energy relationship between a stop and local interference, we calculate the local SNR. For each stop, the local SNR is computed with the whole burst part and 30 ms of the closure. The E_M within a local SNR ranges is shown in Fig. 4. The last data point is the E_M for stops whose local SNRs are larger than 25 dB. Other points correspond to local SNRs with 5 dB increments from 0 dB to 25 dB. When the local SNR is higher than 5 dB, E_M is around 10%. Note that due to the small amount of data within these local SNR ranges, the value of E_M fluctuates around 10%. As local SNR decreases to 0 dB, E_M increases to above 35%

Overall SNR (dB)	E_M(%)	E_F(%)
30	3.3	1.9
20	10.3	5.9
10	33.0	14.6
0	75.5	23.6

Table 1. E_M and E_F

To evaluate the performance of grouping, the speech resynthesized from an ideal binary mask is used as the ground truth for target speech (see [4]). The ideal binary mask is constructed by assigning 1 to a T-F unit where speech before

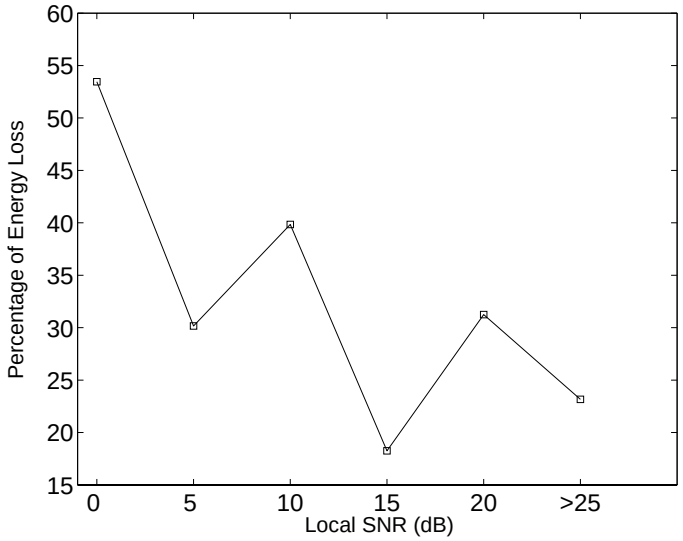

Figure 5. The percentage of energy loss with respect to local SNR

mixing is stronger than interference and 0 otherwise. The use of ideal masks is supported by the auditory masking phenomenon: within a critical band, a weaker signal is masked by a stronger one [6]. In addition, an ideal mask yields excellent recognition performance. Let $O_1(t)$ denote the stop signal resynthesized from the ideal binary mask, and $O_2(t)$ the separated stop signal. Let $e_1(t)$ be the signal present in $O_1(t)$ but missing from $O_2(t)$, and $e_2(t)$ the signal present in $O_2(t)$ but missing from the speech resynthesized from the ideal binary mask. The percentage of energy loss, P_{EL}, and that of noise residue, P_{NR}, are calculated as follows [4]:

$$P_{EL} = \sum_t e_1^2(t) \Big/ \sum_t O_1^2(t) , \tag{10}$$

$$P_{NR} = \sum_t e_2^2(t) \Big/ \sum_t O_2^2(t) . \tag{11}$$

Overall SNR (dB)	P_{EL} (%)	P_{NR} (%)
30	14.61	0.07
20	21.70	0.86
10	35.00	6.42
0	76.67	14.42

Table 2. Average P_{EL} and P_{NR}

Average P_{EL} and P_{NR} at different overall SNR levels are shown in Table 2. The system performs well when SNR is high. As SNR decreases, P_{EL} increases significantly to 77% while P_{NR} increases to 14%. The average P_{EL} for stops with respect to local SNRs is shown in Fig. 5. P_{EL} is around 30% when the local SNR is higher than 5dB. It increases to more than 55% as the local SNR decreases to 0 dB. Two types of error account for the energy loss: missing stops and signals of detected stops that are not grouped into the segregated speech. The first one gives 100% energy loss for each missing stop, while the second one gives approximately 20% energy loss for each detected stop. Note that the plot in fig. 5 has a similar pattern as that in fig. 4, which indicates the relationship between percentage of missing stops and percentage of energy loss.

5. DISCUSSION

We have proposed a method to separate stop consonants, which employs onset as a major ASA cue. This method is able to detect stops and group their frequency components from interfering signals. Onset provides important information for speech segregation, which may be a key to separate unvoiced speech. The onset cue has been studied in previous systems, e.g. [3], but its utility has not been demonstrated. Our approach, i.e., onset detection, feature-based classification, and subsequent grouping, provides a general way to utilize onset information for unvoiced speech. With a comprehensive training, our system may also be adapted

to deal with fricatives and affricates. We include 18 natural sounds in the training, which are far from sufficient to account for general interference since the variety of natural interference is potentially very large, if not infinite. However, the number of most frequently occurring intrusions in a specific environment may be limited. In this case, our system could be trained in a more focused manner more adaptively to the particular environment and hence could perform better.

6. ACKNOWLEDGEMENT

This research was supported in part by an NSF grant (IIS-0081058) and an AFOSR grant (F49620-01-1-0027). A preliminary version has occurred in the proceedings of ICASSP'03.

7. REFERENCES

[1] A. M. Ali, J. Van der Spiegel, and P. Mueller, "Acoustic-phonetic features for the automatic classification of stop consonants," *IEEE. Trans. Speech and Audio Processing*, Vol. 9, 2001, pp. 833-841.

[2] A. S. Bregman, *Auditory scene analysis*, Cambridge, MA: MIT press, 1990.

[3] G. J. Brown and M. P. Cooke, "Computational auditory scene analysis," *Computer Speech and Language*, Vol. 8, 1994, pp. 297-336.

[4] G. Hu and D. L. Wang, "Monaural speech separation," *Proc. of NIPS* 2002.

[5] R. Meddis, "Simulation of auditory-neural transduction: further studies," *J. Acoust. Soc. Am.*, Vol. 83, 1988, pp. 1056-1063.

[6] B. C. J. Moore, *An introduction to the psychology of hearing*, 4th Ed. Academic Press, 1997.

[7] R. D. Patterson, I. Nimmo-Smith, J. Holdsworth, and P. Rice, *APU Report 2341: An efficient auditory filterbank based on the gammatone function*, Cambridge, U.K: Applied Psychology Unit. 1988.

[8] S. J. Russell and P. Norvig, *Artificial intelligence: A modern approach*, 2nd Ed., Prentice Hall, 2003.

[9] D. W. Scott, *Multivariate density estimation*, New York: Wiley & Sons, 1992.

[10] K. N. Stevens, *Acoustic phonetics*, Cambridge, MA: MIT press, 1998.