In: *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing* (ICASSP'03), 2003

Separation of Stop Consonants

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

To extract speech from acoustic interference is a challenging problem. Previous systems based on auditory scene analysis principles deal with voiced speech, but cannot separate unvoiced speech. We propose a novel method to separate stop consonants, which contain significant unvoiced signals, based on their acoustic properties. The method employs onset as the major grouping cue; it first detects stops through onset detection and feature-based Bayesian classification, then groups detected onsets based on onset coincidence. This method is tested with utterances mixed with various types of interference.

I. INTRODUCTION

Speech is often corrupted by acoustic interference. Many applications require a system to extract speech from a mixture in order to improve speech recognition, among other tasks. Currently, no method performs this task well in realistic environments. Previous speech separation efforts utilize harmonicity as the major grouping cue, hence limited to voiced speech [3] [4]. To separate unvoiced speech, other grouping cues must be explored.

In this paper, we address the problem of separating stop consonants from interference. Stops, composed of /t/, /d/, /p/, /b/, /k/, and /g/, constitute a main type of consonants, and they occur frequently in natural speech. A stop generally contains a weak closure and a burst [8]. A closure can be voiced or unvoiced, while a burst is mainly unvoiced and cannot be separated based on harmonicity. The waveform of /g/ and its spectrogram are shown in Fig. 1(a) and 1(b). Our objective is to find the time-frequency regions where stop sounds are dominant, i.e., they are stronger than interference, and group these regions into a target utterance. Because the closure contains little information and is vulnerable to interference, our main strategy for separating stops is to identify dominant stop bursts. Since acoustic properties of dominant bursts are resistant to interference, we first detect stop bursts and then group them accordingly.

At the onset of a stop burst, a significant intensity increase happens across a wide frequency range (see Fig. 1(b)). Therefore, we identify stop bursts by detecting their onsets. Note that onset is an important cue for auditory scene analysis [2]. To detect the onset of a stop burst, an acoustic mixture is first analyzed by an auditory filterbank. Then an onset detector identifies stop candidates by detecting local onsets in a filter channel and integrating across all the channels. Because signals other than stop consonants may also contain portions that are burst-like, these stop candidates are further classified based on three distinctive features: auditory spectrum, relative intensity, and intensity decay time. A Bayesian decision rule is applied for the classification task. Prior probabilities for Bayesian classification are obtained from a training dataset, which contains 100 clean utterances from the TIMIT database for stops, and 18 other natural sounds for interference. Finally, a detected stop burst is recovered by grouping signals starting at its onset. The above method is incorporated into a previous speech segregation system [4], and the resulting system can separate both voiced speech and stop consonants.

This paper is organized as follows. Onset detection and stop classification are described in Sect. II and Sect. III. The results of classification and grouping are given in Sect. IV. Sect. V gives a brief discussion.

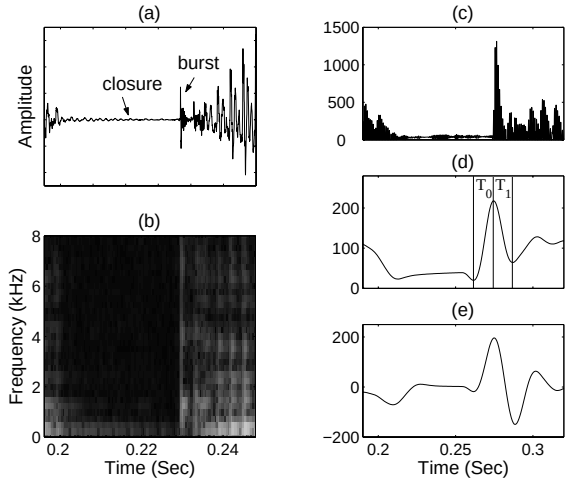

Figure 1. Waveform (a) and spectrogram (b) of /g/ from the word "good"; the corresponding neural firing rate (c), average firing rate (d) and its first-derivative (e) in a channel centered at 1 *k*Hz.

II. ONSET DETECTION

The input signal is sampled at 16 *k*Hz and normalized around 80 dB sound pressure level. It is analyzed by an auditory filterbank, which contains 150 gammatone filters [7] from 80 Hz to 7 *k*Hz, and subsequent neural transduction [5]. The output, in the form of auditory nerve firing rate, is decomposed into 20 ms frames with 10 ms frame shift.

For each channel, possible stop bursts are detected from auditory nerve activity. The activity increases rapidly to a significant level and then decreases slowly to a steady state when the input has a fast intensity increase. Fig. 1(c) shows the auditory nerve firing rate for the input in Fig. 1(a) from the channel centered at 1 *k*Hz. A sudden and significant increase happens at the onset of the burst. We detect onsets by taking the first-derivative of average firing rate, which gives prominent peaks at onset points. This is illustrated in Fig. 1(e). The corresponding average firing rate is shown in Fig. 1(d), which is obtained by lowpassing the auditoy nerve firing rate with a filter (transition band [30 Hz, 80 Hz], passband ripple 0.1, and stopband ripple 0.02). The derivative at time t in channel c, $d(t, c)$, is approximated as follows:

$$d(t,c) = r(t,c) - r(t-\tau,c)\ . \qquad (1)$$

Here, $r(t, c)$ is the average firing rate at time t in channel c, and τ = 14.375 ms, the average of the firing rate rise time of the stops in the training data. The firing rate rise time of a stop is the duration from the corresponding local maximum of the average firing rate to the preceding local minimum (T_0 in Fig. 1(d)). Since the derivative corresponding to onsets is generally greater than the difference between the average steady-state firing rate and the spontaneous firing rate (see [5] for more details), peaks above this difference are marked as channel onsets.

At each frame, our onset detector counts the number of channels containing channel onsets. For stops in the training data, except for a few weak stops, they trigger onsets in at least 10 channels. Therefore, at those frames where 10 or more channels have onsets, the detector identifies a stop candidate, positioned at the local maximum of input signal within the corresponding frame.

III. STOP CLASSIFICATION

Since detected onset candidates may correspond to sources other than stop consonants, we perform stop burst classification based on auditory-acoustic features. Let H_0 denote a hypothesis that a candidate is a stop burst, and H_1 otherwise. Let $\boldsymbol{k}$ be the feature vector for a stop candidate. The likelihood ratio is:

$$L(\boldsymbol{k}) = p(H_0 \mid \boldsymbol{k}) / p(H_1 \mid \boldsymbol{k}) \qquad (2)$$

Here $p(H_j \mid \boldsymbol{k})$ is the posterior probability of H_j given $\boldsymbol{k}$, for j = 0, 1. According to the Bayesian decision rule, the candidate is classified as a stop if and only if $L(\boldsymbol{k})$ is greater than 1. To obtain good classification, one needs to choose appropriate features. Previous research suggests that the following features are important for stop identification: formant transitions, burst spectrum, burst amplitude, durations, and voicing (see [1] for example). Since our main goal is to separate onsets from interference, we choose distinctive features that are robust to acoustic interference.

Each stop has a certain articulatory gesture, which gives unique spectral characteristics [8]. Therefore, we use an auditory spectrum, S_A, as one feature. S_A is obtained as follows:

$$S_A = (r(t_m,1), r(t_m,2), \cdots, r(t_m,150)) / \sqrt{\sum_{c=1}^{150} r^2(t_m,c)}\ . \qquad (3)$$

Here t_m is the location of a stop candidate. We call S_A auditory spectrum since it comes from the output of the auditory filterbank. For each stop phoneme, the average auditory spectrum, obtained from the training data, is shown in Fig. 2. Note that phonemes with the same place of articulation have similar average auditory spectra. We use these averages as templates for stop phonemes. For a stop candidate, the cross-correlation between its auditory spectrum and each template measures the similarity. Among the six cross-correlations, the largest one, denoted by k_S, provides one feature for classification.

The intensity of a stop burst is related to the intensity of neighboring voiced speech [8], while the intensity of interference is generally independent with speech. Therefore, relative intensity of a candidate compared with neighboring voiced speech, denoted by k_I, provides another feature. k_I is obtained as follows:

$$k_I = 10\log_{10}[I_V / I(t_m)], \qquad (4)$$

$$I(t) = \sum_{s=-T}^{T} M^2(t+s) / 2T \qquad (5)$$

Here, $M(t)$ is the input signal, I_V is the average intensity of the input signal in the nearest voiced portion, and T = 1.25 ms.

The intensity of a stop burst drops quickly [8], while a candidate representing interference generally does not. Our third features measures the intensity decay time of a candidate, k_D, which is obtained as follows. Consider the following sequence: {…, I_{-2}, I_{-1}, I_0, I_1, I_2, …}, where $I_n = I(t_m + n\Delta t)$ and Δt = 0.3125 ms, corresponding to 5 samples. Let I_{n1} be the first element after I_0 that is smaller than $0.9I_0$, and I_{n2} the last one before I_0 that is smaller than $0.9I_0$. Then k_D is set to $n1 - n2$.

We use these three features for classification, i.e., $\boldsymbol{k} = (k_S, k_I, k_D)$. For simplicity, we assume that k_S, k_I, and k_D are independent given H_j, for j = 0, 1. This assumption is generally true for interference and a good approximation for stop bursts. Applying Bayesian formula, we have

$$L(\boldsymbol{k}) = \frac{p(k_S \mid H_0)\, p(k_I \mid H_0) P(k_D \mid H_0)\, p(H_0)}{p(k_S \mid H_1)\, p(k_I \mid H_1) P(k_D \mid H_1)\, p(H_1)}\ . \qquad (6)$$

The transition between a stop burst and the following voiced phoneme provides useful information and could be used as another feature. However, it is closely correlated with burst spectrum, hence violating the independence assumption. Also, accurate transition is difficult to obtain from a mixture. Therefore, it is not included.

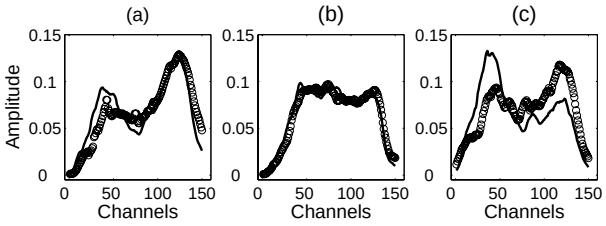

Figure 2. The average auditory spectrum of stop consonants. (a) Circle: /t/; line: /d/. (b) Circle: /p/; line: /b/. (c) Circle: /k/; line: /g/.

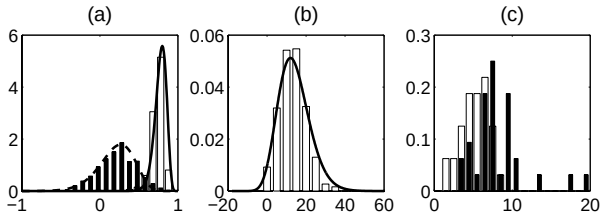

Figure 3. (a) White bar: the histogram of k_S for stops; black bar: the histogram of k_S for interference candidates; solid line: estimated $p(k_S|H_0)$, dash line: estimated $p(k_S|H_1)$. (b) White bar: the histogram of k_I for stops; solid line: estimated $p(k_I|H_0)$. (c) White bar: the histogram of k_D for stops; black bar: the histogram of k_D for interference candidates.

Prior distributions in (6) are obtained from the training data. $p(H_0)$ and $p(H_1)$ are obtained by comparing the average number of candidates from interference and that of stops within the same period of time. From the training date, we have $p(H_0)/p(H_1) = 0.1183$. The histograms of k_C, k_R, and k_T for stop bursts, and those of k_C and k_T for interference in the training data are shown in Fig. 3. Since k_T is discretely distributed, the probabilities from the histograms are directly used in (6). We approximate $p(k_S|H_0)$, $p(k_S|H_1)$, and $p(k_I|H_0)$ each with a Chi-square distribution:

$$p(k_S \mid H_0) = a_0 g_\chi(80 - 80c, 18) \quad (7)$$

$$p(k_S \mid H_1) = a_1 g_\chi(30 - 30c, 24) \quad (8)$$

$$p(k_I \mid H_0) = b_0 g_\chi(18 + c, 32), \quad (9)$$

where

$$g_\chi(x,n) = x^{n/2-1}e^{-x/2}u(x)/[2^{n/2}\Gamma(n/2)],$$
$$\Gamma(\alpha) = \int_0^\infty x^{\alpha-1}e^{-x}dx \quad . \quad (10)$$

Here, a_0, a_1, b_0 are constants for normalizing overall probability. These distributions are shown in Fig. 3. For $p(k_I|H_1)$, we simply use a uniform distribution from –20 dB to 60 dB according to the following considerations. First, the onset detector is generally will not sensitive to signals that are more than 60 dB below voiced speech. Second, signals that are more than 20 dB above voiced speech are unlikely to be stop.

IV. RESULTS

A detected stop burst is separated from acoustic interference as follows. Because signal starting at the onset of a stop burst is likely to be part of the burst, the channels with onsets less than 5 ms away from the stop candidate are selected. Since stop bursts are short, in each selected channel, a burst-dominant region is chosen from the corresponding local maximum to the next local minimum of the average firing rate (T_1 in Fig. 1(d)); a local minimum of the average firing rate usually corresponds to an offset of a sound. This region generally contains the major part of the burst. Time-frequency (T-F) units that are mostly occupied by this region 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 binary mask is used to resynthesize a target utterance. It retains the acoustic energy from the mixture corresponding to 1's and rejects that corresponding to 0's (see [3] for more details).

This method is tested with 10 utterances from the TIMIT database mixed with the following 10 interference: white noise, pink noise, airplane noise, car noise, factory noise, noise burst, clicks, bar noise, a firework show, and rain. None of the utterances are used in the training. To evaluate the performance of stop detection, let E_M be the percentage of stops that are missing, and E_F be the percentage of interfering signals wrongly detected as stops. Note that the overall error rate is $p(H_0)E_M + p(H_1)E_F$. E_M and E_F at different overall SNR levels are shown in Table 1. E_M increases significantly as SNR decreases since more stops are corrupted by stronger interference, while E_F only increases slightly. Note that the Bayesian classifier is designed to distinguish 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. Note that the goal of speech separation is to remove interference.

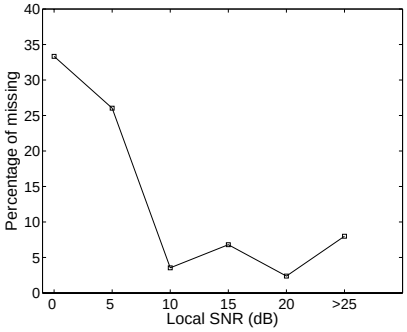

Figure 4. The percentage of missing stops with respect to local SNR

Table 1. E_M and E_F

Overall SNR (dB)	E_M(%)	E_F(%)
30	8.6	0.6
20	24.8	1.4
10	62.6	2.8
0	84.0	6.8

The overall SNR shows the energy ratio between voiced speech and interference. To get more insight into the performance related to the energy relationship between a stop and local interference, we calculate the local SNR, which includes a 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. E_M is smaller than 10% when the local SNR is higher than 10 dB. It drops to 40% as local SNR decreases to 0 dB.

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 mixing is stronger than interference and 0 to 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

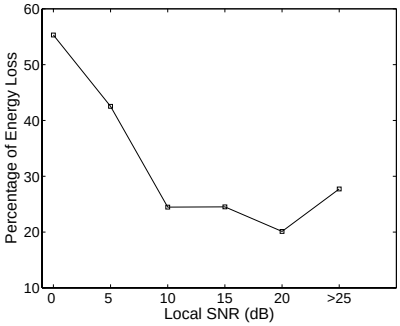

Figure 5. The percentage of energy loss with respect to local SNR

mask. The percentage of energy loss, P_{EL}, and that of noise residue, P_{NR}, are calculated as follows [4]:

$$P_{EL} = \sum_t e_1^2(t) \Big/ \sum_t O_1^2(t), \quad (11)$$

$$P_{NR} = \sum_t e_2^2(t) \Big/ \sum_t O_2^2(t) \quad (12)$$

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 85% while P_{NR} increases to around 10%. The average P_{EL} for stops with respect to local SNRs is shown in Fig. 5. P_{EL} is below 30% when the local SNR is higher than 10 dB. It increases to 55% as the local SNR decreases to 0 dB.

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

Overall SNR (dB)	P_{EL} (%)	P_{NR} (%)
30	28.01	0.04
20	41.56	0.81
10	70.68	2.81
0	84.79	9.62

To further illustrate the performance of our method, we incorporate it in a previous voiced speech separation system [4] in order to separate utterances containing both voiced speech and stop consonants. More specifically, the previous system marks speech dominant T-F units in the voiced part, and the proposed method marks those units in the stop bursts. These marked units are combined to generate a binary mask for resynthesis. Fig. 6 illustrates the separated speech from a mixture containing a male utterance, "A good morrow to you, my boy", and rain at 10 dB. Among four stops, two are separated. For this mixture, P_{EL} is 53.37%, and P_{NR} is 12.76%.

V. DISCUSSION

Through onset detection and feature-based Bayesian classification, we are able to detect stops and separate most of them from interfering signals. As a major ASA cue, onset provides important information for separating unvoiced speech and dealing with reverberation. The onset cue has been studied in some previous systems, e.g. [3], its utility has not been demonstrated. Our method for stop separation, i.e., onset detection, feature-based classification, and subsequent grouping, provides a general approach to utilize onset information for speech separation. In fact, the onset detector can detect onsets of other phonemes as well as stops. To deal with general speech,

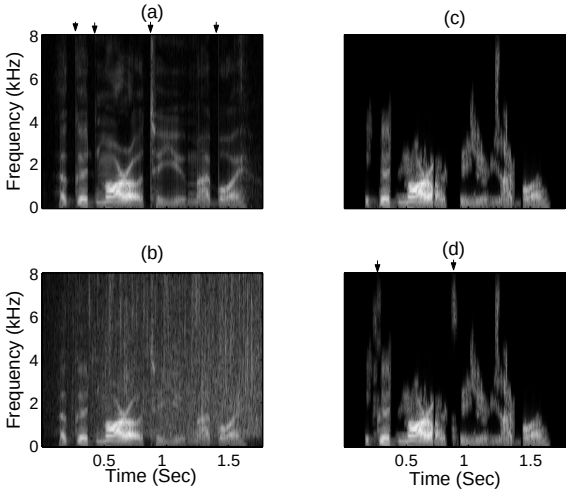

Figure 6. (a) The spectrogram of a male utterance. Stops are marked by arrows. (b) The spectrogram of this utterance mixed with rain sound at 10 dB SNR. (c) The separated speech from the system of [4]. (d) The separated speech from the extended system. The recovered stops are marked by arrows.

more comprehensive training would be needed to build a classifier that is capable of removing onsets from interference and classifying onsets caused by different phonemes. We plan to explore these issues in future research.

ACKNOWLEDGEMENT. This research was supported in part by an NSF grant (IIS-0081058) and an AFOSR grant (F49620-01-1-0027).

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 based on pitch tracking and amplitude modulation," *ICASSP* 2002, pp. 553-556.

[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: Applied Psychology Unit. 1988.

[8] K. N. Stevens, *Acoustic phonetics*, MIT press: Cambridge, Massachusetts, 1998.