Features for Automatic Detection of Voice Bars in Continuous Speech

Dhananjaya N[1], *S. Rajendran*[2] *and B. Yegnanarayana*[2]

[1]Indian Institute of Technology Madras, India
[2]International Institute of Information Technology Hyderabad, India
dhanu@cs.iitm.ernet.in, su.rajendran@gmail.com, yegna@iiit.ac.in

Abstract

In this paper we propose features for automatic detection of voice bar, which is an essential component of voiced stop consonants, in continuous speech. The acoustic-phonetic and production based knowledge such as, the presence of voicing, low strength of excitation compared to other voiced phones and a predominant low-frequency spectral energy, are mapped onto a set of acoustic features that can be automatically extracted from the signal. The usefulness of the proposed features in the detection of voice bars is studied using a knowledge-based as well as a neural network based approach. The performance of the proposed features and approaches is studied on phones from databases of two languages, namely English and Hindi.

Index Terms: voice bars, voiced stop-consonants, acoustic-phonetic, residual-to-signal ratio, normalized error, zero-frequency resonator, knowledge based approach, neural network based approach.

1. Introduction

Voice bar refers to the voiced region of the acoustic waveform corresponding to the phonation when the oral and nasal cavities are completely closed. It is the acoustic manifestation of the sound radiated through the pharyngeal wall. The voice bar is the essential component of voiced stop consonant and it corresponds to the silent event of voiceless stop consonant. The closure in the oral cavity takes place anywhere between the upper lip and pharynx depending upon the type of the stop consonant. Detection of voice bar from continuous speech not only helps in identifying the voiced stop consonant concerned, but is also useful for automatic segmentation and labeling of speech corpora for speech synthesis as well. However, detection of voice bar in continuous speech is considered as a difficult problem due to the poor acoustic signal strength. Several studies made on the classification of stop-consonants [1] assume that the segmentation or detection of these stop-consonants is already done. Accurate detection of stop-consonants in continuous speech is essential in order to use these classification strategies.

The organization of the paper is as follows: In Section 2 the acoustic features proposed for the detection of voice bars are described. Section 3 describes the methods employed for detection of voice bars in continuous speech. The dataset used for the experiments and the performance of the voice bar detection task are discussed in Section 4. Summary and conclusions are given in Section 5.

2. Features for detection of voice bars

Many studies have been made in identifying the acoustic cues for the segmentation and labeling of speech [2, 3, 4]. The main issue with these studies is that the data used for analysing the acoustic cues are from well-articulated isolated utterances of the phonetic units. But in continuous speech, these acoustic features are not properly manifested due to aspects like accent, emotion and speaking rate of the speaker, along with the coarticulation of neighbouring sounds depending on the context in which the phone occurs. Also, accurate measurement of these acoustic features from the speech signal is an important issue. In this section, we propose a set of acoustic features for automatically identifying regions of voice bar, in continuous speech.

2.1. Residual to signal ratio

The residual to signal ratio (RSR) or the normalized linear prediction (LP) error is computed as, $v_{rsr}[n] = e_r[n]/e_s[n]$, where $e_r[n] = 1/N\sum_{i=-N/2}^{N/2} r^2[n+i]$ is the short term energy of the linear prediction residual signal obtained by inverse filtering the speech signal. Similarly, $e_s[n]$ is the short term energy of the speech signal. A 10^{th} order short-term (20 ms frame size and 10 ms frame shift) LP analysis is performed to compute the residual signal. The inverse filtering removes most of the signal energy from the voiced regions as compared to nonvoiced (unvoiced and silence) regions of speech, which typically have uncorrelated speech samples. This results in a low RSR value in the voiced regions compared to nonvoiced regions, as can be seen in Figure 1(b). Any channel related correlations are removed by preemphasizing the speech signal before LP analysis, using a simple difference operation. This feature primarily helps identifying voiced regions from nonvoiced regions.

2.2. Low- to high-order residual energy ratio

The low- to high-order residual (LHR) energy ratio is computed as, $v_{lhr}[n] = e_{r_1}[n]/e_{r_{10}}[n]$, where $e_{r_1}[n]$ and $e_{r_{10}}[n]$ are the short-term energies of LP residual signals of order 1 and 10, respectively. A 1^{st} order LP inverse filtering removes most of the signal energy from the voice bar regions which have a predominantly low frequency content. At the same time, only a small portion of the energy is removed from other voiced regions of the speech signal. In comparison, a 10^{th} order LP inverse filtering removes as much energy in the voice bar regions, while it removes a significant amount of energy from the other voiced regions. This can be seen from the 1^{st} and 10^{th} order normalized error signals shown in Figure 1(c). Hence a ratio of the normalized errors, shown in Figure 1(d), can be used as a feature for discriminating voice bars from other voiced regions.

2.3. Zero-frequency resonator signal strength

A zero-frequency resonator (ZFR) is a linear time-invariant all-pole system with two real poles on the positive real axis of the z-plane[5]. The proximity of the poles to the unit circle determine the bandwidth of the resonator. The speech signal

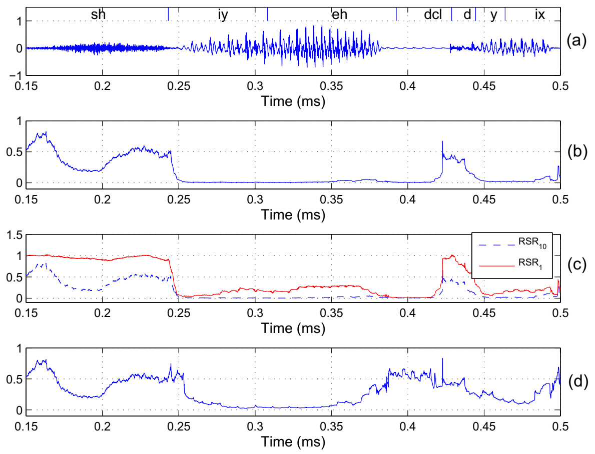

Figure 1: *Acoustic features based on normalized error. (a) The speech waveform for an utterance "She had you...". Manually marked phoneme labels are given above the signal. The label 'dcl' around 0.4 sec corresponds to the voice bar region (closure region of the voiced stop consonant /d/). (b) The* 10^{th} *order RSR signal. (c) RSR signals for LP orders 1 (solid) and 10 (dashed). (d) Reciprocal of the LHR signal.*

passed through a zero-frequency resonator predominantly contains the low frequency content. The zero-frequency resonator introduces a cumulative DC bias which is removed by subtracting the local mean computed using a moving average window of size 10 ms. The ZFR signal and its short-term (20 ms) energy are shown in Figure 2. It can be seen that the ZFR signal energy is a good evidence for discriminating voiced regions from the nonvoiced.

2.4. ZFR signal to speech energy ratio

The zero-frequency resonator signal to speech energy ratio is computed as $v_{zsr}[n] = e_z[n]/e_s[n]$, where $e_z[n]$ and $e_s[n]$ are the short-term (20 ms) energies of the zero-frequency resonator signal and the speech signal, respectively. The speech and the ZFR signal are normalized to an overall root-mean-square (RMS) value of unity, before the computation of short-term energies. The normalized speech and ZFR signals are shown in Figures 2(a) and (b), respectively. It can be seen that the relative amplitude of the voice bar region with respect to the adjacent vowel regions is more in the ZFR signal as compared to that in the speech signal. This is due to the fact that the ZFR allows most of the energy in voice bar regions while allowing only a part of the energy in other voiced regions. Hence the ZFR signal to speech energy ratio is higher for voice bar regions compared to other voiced regions, as can be seen in Figure 2(d).

2.5. Strength of excitation source signal

A 10^{th} order LP residual signal which is void of most of the vocal tract system characteristics is used as an estimate of the excitation source signal. The envelop of the LP residual signal is computed as the magnitude of the complex analytic signal obtained by Hilbert transform of the residual signal. The residual signal has large errors around the glottal closure instants (GCIs), which appear as peaks in the Hilbert envelop signal. The amplitude of the Hilbert envelop signal at the peak locations correspond to the rate at which the vocal folds close, and hence

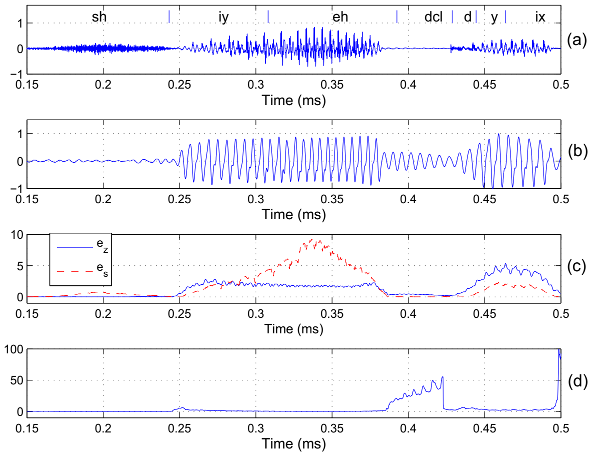

Figure 2: *Acoustic features from the zero-frequency resonator signal. (a) The speech signal, (b) ZFR signal, (c) short-term energies of the ZFR (solid) and speech (dashed) signals, and (d) ZFR signal to speech energy ratio.*

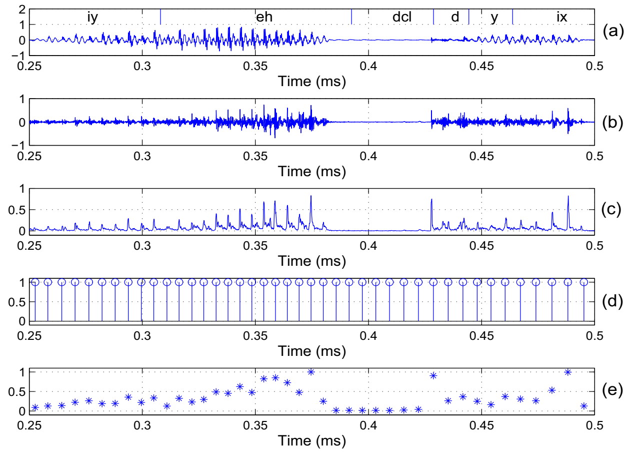

Figure 3: *Excitation strength based feature. (a) Speech signal, (b) LP residual of order 10, (c) Hilbert envelop of the residual signal, (d) Glottal closure instants, and (e) strength of excitation source.*

is taken as a measure of the strength of the excitation source signal. The peaks in the envelop signal are located by picking the positive to negative zero-crossings of the short-term (10 ms) phase slope function[6]. The speech, residual and Hilbert envelop signals along with the GCIs and the excitation strength are shown in Figure 3. The strength of excitation for voice bars is typically low compared to other voiced sounds, and hence is a useful feature to discriminate voice bars from other voiced sounds in continuous speech.

2.6. Dominant resonance frequency

The dominant resonance frequency (DRF) is measured by picking the largest peak in the numerator group-delay (NGD) spectrum [7]. The NGD spectrum is used as it resolves the spectral peaks better than the magnitude spectrum, for short (less than a pitch period) segments of speech [7]. For a given signal $x[n]$, the NGD is computed as $\tau(w) = (X_R(w) * Y_R(w) + X_I(w) * Y_I(w))$, where $X(w)$ and $Y(w)$ denote the discrete

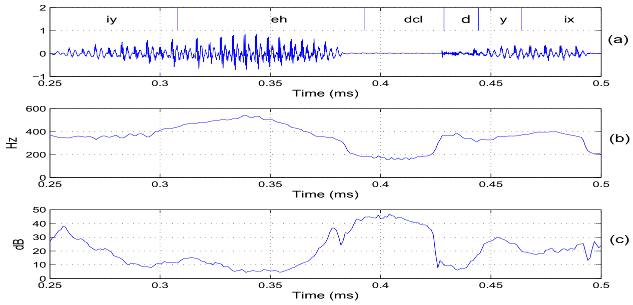

Figure 4: *Acoustic features based on dominant resonance frequency. (a) Speech signal, (b) short-term DRF, and (c) short-term DRS.*

time Fourier transform of $x[n]$ and $y[n] = nx[n]$. The subscripts $_R$ and $_I$ denote the real and imaginary parts of the complex spectrum. The DRF computed for every 6 ms window with 1 ms shift is shown in Figure 4(b). Due to the predominantly low spectral energy in voice bars, the dominant resonance frequency is low for voice bars compared to most of the other voiced sounds.

2.7. Normalized dominant resonance strength

The strength of the dominant resonance frequency is measured as the magnitude of the numerator group-delay spectrum at the DRF. It is normalized by the strength of the second most dominant peak in the NGD spectrum. The normalized dominant resonance strength (DRS) is computed as $v_{drs} = 20 * \log_{10}(\tau_n(f_1)/\tau_n(f_2))$, where $\tau_n(f)$ is the NGD, and f_1 and f_2 are the first and second most dominant frequencies in the NGD. The normalized DRS is higher for voice bars compared to most of the other voiced sounds as can be seen in Figure 4(c).

3. Automatic detection of voice bars

Two different approaches - a knowledge based approach and a neural network based approach - are explored to study the usefulness of the proposed features in the detection of voice bars in continuous speech. The knowledge based approach (KBA) gives a better insight into the speech production mechanism and the acoustic signal. But the main problem with this approach is the setting of thresholds. The neural network approach (NNA) avoids this problem, but requires manually and accurately labeled data for training the neural network model.

3.1. Knowledge based approach

A hierarchical evidence-based classification strategy is employed using the empirical knowledge acquired by manual analysis of the acoustic signal and the features. The algorithm employed for the detection of voice bars is as follows:

1. The voiced-nonvoiced decision is arrived at using three of the features described in Section 2, namely $v_{rsr}[n]$, the 10^{th} order RSR signal, $v_{zsr}[n]$, ZFR signal to speech energy ratio, and $v_{zfr}[n]$, the zero frequency resonator signal strength. The binary voiced-nonvoiced signal is computed as,

$$d_{vnv}[n] = \begin{cases} 1 \ if \ (\frac{y_{rsr}[n]+y_{zsr}[n]+y_{zfr}[n]}{3}) > 0.5 \\ \\ 0 \ otherwise, \end{cases} \tag{1}$$

where $y_{rsr}[n] = 1 - e^{(-10*v_{rsr}[n])}$, $y_{zsr}[n] = 1 - e^{(-v_{zsr}[n])}$ and $y_{zfr}[n] = 1 - e^{(-10*v_{zfr}[n])}$.

2. The first level of evidence for discriminating voice bars from other voiced sounds is obtained based on the ZFR signal to speech energy ratio $v_{zsr}[n]$, the LHR energy ratio $v_{lhr}[n]$, and the voicing decision $d_{vnv}[n]$ obtained in the previous step. It is computed as,

$$d_{vb_1}[n] = \begin{cases} 1 \ if \ \{(d_{vnv}[n]) \ \& \ (y_{zsr}[n] > 0.99) \\ \qquad\qquad\qquad\quad \& \ (y_{lhr}[n] < 0.99)\} \\ \\ 0 \ otherwise, \end{cases} \tag{2}$$

where $y_{lhr}[n] = 1 - e^{-10*(v_{lhr}[n]-2))}$.

3. The final decision on the locations of voice bars is made by validating d_{vb_1} using the excitation strength and dominant resonance evidences, and is computed as,

$$d_{vb}[n] = \begin{cases} 1 \ if \ \{(d_{vb_1}[n]) \ \& \ (v_{es}[n]/v_{mes} < 0.1) \\ \quad \& \ (v_{drf}[n] < 300) \ \& \ (v_{drs}[n] > 25)\} \\ \\ 0 \ otherwise. \end{cases} \tag{3}$$

The excitation strength $v_{es}[n]$ is required to be less than 10% of the maximum excitation strength v_{mes} computed over the entire signal. The dominant resonance frequency should be less than 300 Hz and its strength should be at least 25 db more compared to the next dominant frequency.

Figure 5 shows a portion of a speech utterance along with the various evidence and decision plots. All the thresholds are chosen empirically and in a conservative manner so as not to miss a genuine voice bar even if a few false alarms are allowed.

3.2. Neural network based approach

A multilayered feedforward neural network (MLFFNN) classifier is used to automatically learn the nonlinear decision surface between the voice bars and the rest of the voiced sounds. A four layered MLFFNN neural network model is used. It consists of an input layer with as many linear nodes as the number of features, two hidden layers with nonlinear nodes and an output layer with one nonlinear node for a two class classification. The first hidden layer is used as an expansion layer which helps in nonlinear transformation of the input feature vector into a higher dimension space where the patterns are more linearly separable. A second hidden layer provides for a gradual transformation from a high dimension space to the one-dimension space of the output layer. The activation function used for the nonlinear nodes are $tanh$ functions. Standard backpropagation learning algorithm is used for training the neural network.

4. Experimental results

4.1. Datasets

The performance of the proposed approaches for detecting voice bars in continuous speech is evaluated on datasets of two

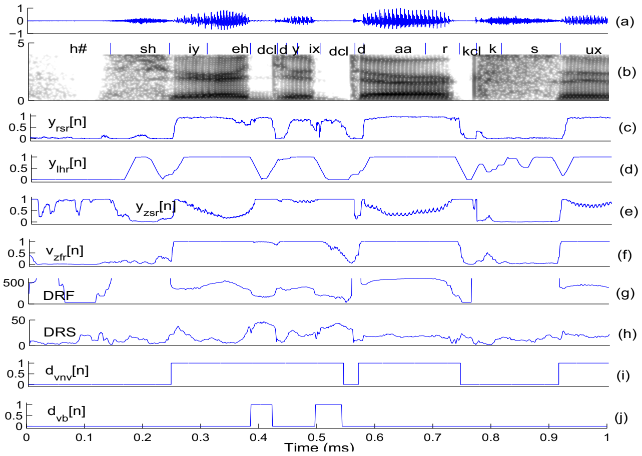

Figure 5: *Features and evidence plots used in knowledge based approach. (a) Speech signal, (b) spectrogram of the speech signal along with manually marked phones, (c) $y_{rsr}[n]$, (d) $y_{lhr}[n]$, (e) $y_{zsr}[n]$, (f) $v_{zfr}[n]$, (g) DRF, (h) DRS, (i) voicing decision, and (j) the final voice bar decision.*

different languages, namely English and Hindi. A subset of data from the standard TIMIT database [8] is used for the studies on English language phones. The data from one of the dialects 'dr1' is used for acquiring the knowledge as well as training the neural network model. The data from another dialect 'dr6' is used for testing. The dialect 'dr1' has 38 speakers (24 male and 14 female) each uttering ten short (3 to 4 seconds) sentences. The dialect 'dr6' has 35 speakers (22 male and 13 female) each uttering ten short (3 to 4 seconds) sentences. A small subset of the Hindi speech database used in synthesis experiments described in [9], is used to study the performance of the proposed features on the phones of Hindi language. The data used is about five minutes of manually marked, single female speaker speech recorded in a quiet room at a sampling rate of 16 kHz. There is no separate data used for knowledge acquisition or for training. The performance is evaluated based on the knowledge acquired or model trained using the data from dialect 'dr1' of the TIMIT dataset.

4.2. Performance evaluation

The performance of the task of detecting voice bars in continuous speech is evaluated in terms of the missed detection rate and false alarm rates. The missed detection rate is computed as $P_m = N_m/N_{vb} * 100\%$, where N_m is the number of missed voice bars out of a total N_{vb} voice bars. The false alarm rate is computed as $P_f = N_f/N_{nvb} * 100\%$, where N_f is the number of nonvoice-bars that are detected as voice bars out of a total number of N_{nvb} nonvoice-bars.

A voice bar is said to be detected correctly if at least 10 ms of the closure period is detected. Similarly, any phone not having a voice bar and with more than 10 ms of its region marked as voice bar is counted as a false alarm. Table 1 gives the performance of the two approaches for detecting voice bars in continuous speech. It is seen that the neural network based methods perform better than the knowledge based methods, due to their ability to automatically learn the thresholds. The network structure used is '7L-14N-5N-1N', where L or N denote a linear or a nonlinear layer, and the preceeding number specifies the number of nodes in the layer. The network structure is determined empirically. Most of the missed voice bars are due to poor articulation of the voiced stop-consonants without any voicing in the closure period. The manual labeling by human listeners is mainly driven by what one expects to hear than what is actually uttered. Most of the false alarms are due to nasals (/m/ and /n/) and semivowels (liquids /l/ and /w/) which have a lot of similarities to the voice bars in terms of the acoustic features used.

Table 1: *Errors in detection of voice bars in continuous speech*

Dataset	Approach	P_m (%)	P_f (%)
TIMIT-dr1	KBA	13.9	11.8
	NNA	11.3	9.1
TIMIT-dr6	KBA	14.8	13.1
	NNA	11.9	10.5
HINDI	KBA	9.2	10.3
	NNA	7.4	7.9

5. Summary and Conclusions

In this paper, the problem of identifying the locations of voice bars in continuous speech was addressed. A set of acoustic features were proposed and evaluated for their performance using two different approaches - knowledge based and neural network based. It was shown that good detection accuracies can be obtained using both the methods. The identification of voice bar regions in continuous speech provide for good anchor points for further segmentation and labeling of speech. We are currently working on using the proposed set of features to detect regions of speech such as frication, burst, aspiration and vowel onset points.

6. References

[1] Ahmed. M. Abdelatty Ali, Jan Van der Spiegel, and Paul Mueller, "Acoustic-phonetic features for the automatic classification of stop consonants," *IEEE Trans. Speech and Audio Processing*, vol. 9, no. 8, pp. 833–841, Nov. 2001.

[2] K. N. Stevens, *Acoustic Phonetics*. Cambridge, Massachusetts, USA: The MIT Press, 1999.

[3] J. Harrington, "Acoustic cues for automatic recognition of English consonants," in *Aspects of speech technology*, M. Jack and J. Laver, Eds. UK: Edinburgh University Press, 1988, pp. 69–143.

[4] S. E. G. Ohman, "Coarticulation in VCV utterances: Spectragraphic measurements," *Journal of the Acoustical Society of America*, vol. 39, pp. 151–168, 1966.

[5] B. Yegnanarayana, K. Sri Rama Murty, and S. Rajendran, "Analysis of stop consonants in Indian languages using excitation source information in speech signal," accepted for publication in *Proc. Speech Analysis and Processing for Knowledge Discovery* to be held in Aalborg, Denmark during June 4-6, 2008.

[6] R. Smits and B. Yegnanarayana, "Determination of instants of significant excitation in speech using group delay function," *IEEE Trans. Speech and Audio Processing*, vol. 3, no. 5, pp. 325–333, Sept. 1995.

[7] Anand Joseph M, Guruprasad S, and B. Yegnanarayana, "Extracting formants from short segments of speech using group delay functions," in *Proc. Int. Conf. Spoken Language Processing (INTERSPEECH)*, Pittsburgh PA, USA, Sept. 2006, pp. 1009–1012.

[8] John S. Garofolo, et al., *TIMIT Acoustic-Phonetic Continuous Speech Corpus*, Linguistic Data Consortium, Philadelphia, USA, 1993.

[9] S. P. Kishore and A. W. Black, "Unit size in unit selection speech synthesis," in *Proc. Eurospeech*, Geneva, Switzerland, Sept. 2003, pp. 1317–1320.