Speaker Segmentation Based on Subsegmental Features and Neural Network Models

N. Dhananjaya, S. Guruprasad, and B. Yegnanarayana

Speech and Vision Laboratory
Department of Computer Science and Engineering
Indian Institute of Technology Madras, Chennai-600 036, India
{dhanu,guru,yegna}@cs.iitm.ernet.in

Abstract. In this paper, we propose an alternate approach for detecting speaker changes in a multispeaker speech signal. Current approaches for speaker segmentation employ features based on characteristics of the vocal tract system and they rely on the dissimilarity between the distributions of two sets of feature vectors. This statistical approach to a point phenomenon (speaker change) fails when the given conversation involves short speaker turns (< 5 s duration). The excitation source signal plays an important role in characterizing a speaker's voice. We use autoassociative neural network (AANN) models to capture the characteristics of the excitation source that are present in the linear prediction (LP) residual of speech signal. The AANN models are then used to detect the speaker changes. Results show that excitation source features provide better evidence for speaker segmentation as compared to vocal tract features.

1 Introduction

Given a multispeaker speech signal, the objective of speaker segmentation is to locate the instants at which a speaker change occurs. Speaker segmentation is an important preprocessing task for applications like speech recognition, audio indexing and 2-speaker detection. Human beings perceive speaker characteristics at different (signal) levels, which, based on the duration of analysis, can be grouped into segmental (10-50 ms), subsegmental (1-5 ms) and suprasegmental (> 100 ms) features. Most of the current methods for speaker segmentation use the distribution of short-time (segmental) spectral features relating to the vocal tract system, estimated over five or more seconds of speech data, to detect speaker changes. However, these methods cannot resolve speaker changes over shorter durations of data (< 5 s), owing to their dependence on the statistical distribution of the spectral features.

The objective of this study is to explore features present in the source of excitation, to the vocal tract system, for speaker segmentation. In section 2, we give a review of the current approaches to speaker segmentation and bring out their limitations in detecting speaker changes due to short (< 5 s) speaker turns. Section 3 describes the use of autoassociative neural network (AANN) models in

N.R. Pal et al. (Eds.): ICONIP 2004, LNCS 3316, pp. 1210–1215, 2004.

characterizing a speaker from the subsegmental features present in the excitation source signal. In section 4 we propose a speaker segmentation algorithm using excitation source features. The performance of the proposed method in speaker segmentation is discussed in section 5. Section 6 summarizes the work and lists a few issues still to be addressed.

2 Need for Alternate Approaches to Speaker Segmentation

Current methods for speaker segmentation use features representing the vocal tract system of a speaker. Two adjacent regions of speech are compared for dissimilarity in the statistical distributions of the feature vectors. Mel-frequency cepstral coefficients (MFCC) or linear prediction cepstral coefficients (LPCC) are used as feature vectors. Some widely used dissimilarity measures include the delta-Bayesian information criterion (dBIC) [1] [2] and Kullback-Leibler distance [2]. In [3], generalized likelihood ratio is used as distance measure to separate out a dominant speaker from other speakers in an air traffic control application. In [2], a multipass algorithm for detecting speaker changes is presented, which uses various window sizes and different dissimilarity measures over different passes. In all these studies, large (> 5 s) speaker turns are hypothesized, while the short turns do not receive attention owing to the application under consideration.

To illustrate the inadequacy of spectral features for speaker change detection, the performance of BIC approach is studied on two types of 2-speaker data, one with long speaker turns and the other with short speaker turns, and is shown in Fig. 1 and Fig. 2, respectively. 19-dimensional weighted LPCCs, obtained from a 12^{th} order LP analysis, are used as feature vectors, and dBIC is used as the dissimilarity measure. It is seen from Fig. 1 that the evidence for speaker change reduces drastically as the window size is reduced, while Fig. 2 illustrates the inability of BIC method in picking the speaker changes with short speaker turns.

3 Speaker Characterization Using Subsegmental Features

Linear prediction (LP) analysis of speech signal gives a reasonable separation of the vocal tract information (LP coefficients) and the excitation source information (LP residual) [4, 5]. If the LP residual of a voiced segment of speech is replaced by a train of impulses separated by one pitch period, and speech is synthesized using the same LP coefficients, it is observed that many of the speaker characteristics are lost. Thus, it is hypothesized that the voiced excitation has significant speaker-specific characteristics. An autoassociative neural network (AANN) model can be used to capture the higher order relations among the samples of the LP residual signal [6]. Blocks of samples of the LP residual (derived over voiced regions) are presented as input to the AANN model. These blocks are presented in a sequence, with a shift of one sample. The blocks are

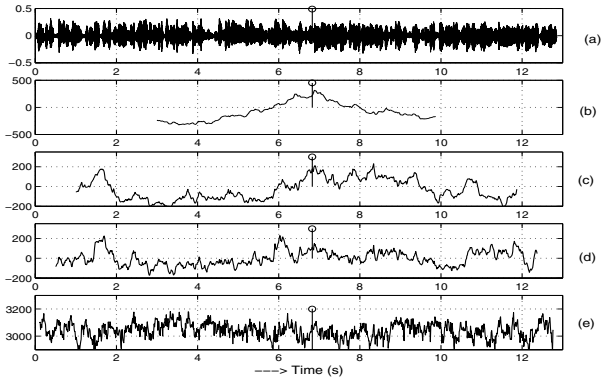

Fig. 1. Case 1: Long (> 5 s) speaker turns. Deteriorating evidence with reducing window sizes. (a) 2-speaker speech signal. (b) to (e) dBIC plots for window sizes of 3 s, 1 s, 0.5 s and 0.1 s respectively. True speaker change is marked by a vertical pole.

typically less than a pitch period in size (subsegmental) and are normalized to unit magnitude before presenting to the AANN model. Once an AANN model is trained with the samples of the LP residual, blocks of samples from a test signal can be presented in a manner similar to the training data. The error between the actual and desired output is obtained, and is converted to a confidence score using the relation, $c = exp(-error)$. The AANN model gives a high confidence scores if the test signal is from the same speaker.

4 Proposed Method for Speaker Segmentation

The algorithm for speaker change detection has two phases, a model building phase and a change detection phase.

Model Building Phase:
An AANN model is trained from approximately 2 sec of contiguous voiced speech which is hypothesized to contain only one speaker. In a casual conversational speech it is not guaranteed that a single random pick of 2 sec data contains only one speaker. In order to circumvent this problem, M (about 10) models are built from M adjacent speech segments of 2 sec, with an overlap of 1 sec. The possibility of at least two pure segments (of a single speaker) is thereby increased. The entire conversation is tested through each of the models to obtain M confidence plots. The cross-correlation coefficients between all possible pairs of confidence plots are computed. N (2 or 4) out of M models are picked which give high correlation coefficient value with each other. The entire process of model building and selection is depicted in Figure 3.

Change Detection Phase:
This phase involves combining evidence from the chosen N confidence plots after model selection. An absolute difference $\triangle\mu$, of average confidence scores from two adjacent window segments (500 ms) is computed to obtain the $\triangle\mu$ plot by

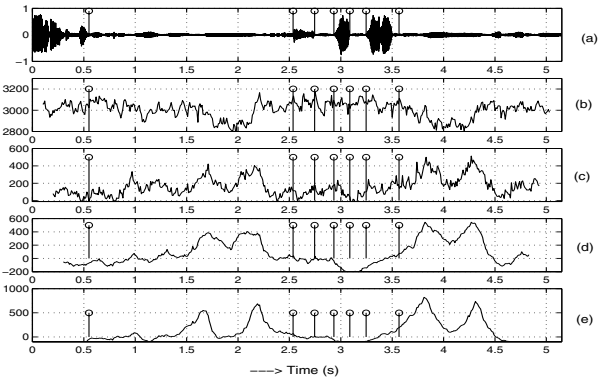

Fig. 2. Case 2: Short (< 5 s) speaker turns. Illustrating lack of evidence for speaker change detection. (a) 2-speaker speech signal. (b) to (e) dBIC plots for window sizes of 0.1 s, 0.2 s, 0.5 s and 1 s respectively. True speaker changes are marked by vertical poles.

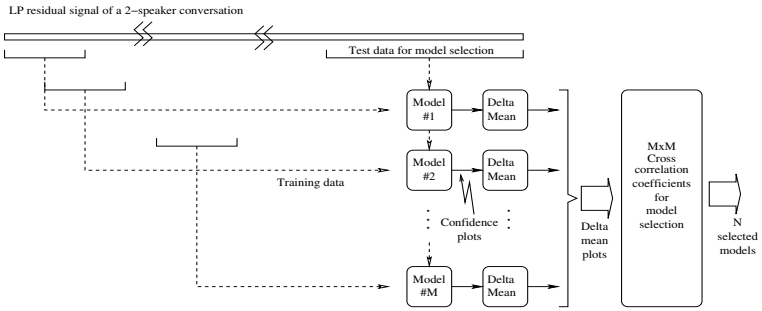

Fig. 3. Model building and selection process.

shifting the pair of windows by 5 ms. Figure 4(b), (c), (d) and (e) show the evidence for the chosen four AANN models. The four evidences are combined using *AND* logic and the result is shown in Figure 4(f). The dBIC plot for the same 2-speaker data, given in Figure 4(g), shows relatively poorer evidence when vocal tract features are used.

5 Performance of the Proposed Approach

Performance Metrics:
The performance of speaker segmentation is evaluated using the false acceptance or alarm rate (FAR) and the missed detection rate (MDR). FAR is the number of false speaker changes, while MDR is the number of missed speaker changes, both expressed as a percentage of the actual number of speaker changes. An ideal system should give an FAR of 0% and an MDR of 0%. The performance of the segmentation is also measured in terms of the segmentation cost function given by, $C_{seg} = 1 - T_c/T_t$,where T_c is the total duration of voiced speech (in

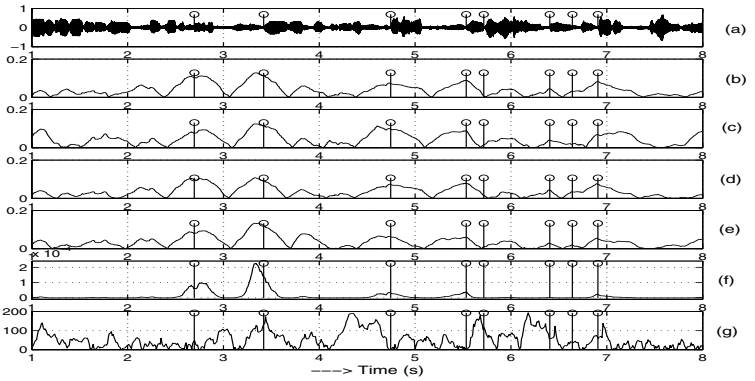

Fig. 4. Combining evidence for speaker change detection. (a) 2-speaker signal with short speaker turns. $\triangle\mu$ plots for (b) model 1, (c) model 2, (d) model 3, (e) model 4, (f) combined evidence and (g) dBIC for vocal tract features. Solid lines indicate the actual speaker change points.

time) correctly segmented and T_t is the total duration of the voiced speech in the conversation. The cost function is normalized by a factor $C_{default}$, to obtain a normalized segmentation cost $C_{norm} = C_{seg}/C_{default}$. $C_{default}$ is the minimum segmentation cost that can be obtained even without processing the conversation (by assigning the entire conversation to either of the speaker). A good system should give a C_{norm} value close to zero, and a value close to one is as good as not processing the conversation.

Data Set for Performance Evaluation:
A total of 10 different 2-speaker conversations each of duration 5 minutes are used to evaluate the performance of speaker segmentation system. The 2-speaker speech signals are casual telephonic conversations and are part of the NIST-2003 database for speaker recognition evaluation [7]. Out of the 10 conversations, 5 are male-male conversations and 5 are female-female conversations. The data set has a total of 1047 actual speaker changes (manually marked). A five layered AANN model with a structure $40L60N12N60N40L$ is used in the experiments and the residual samples are fed to the neural network in blocks of 5 ms. The FAR, MDR and C_{norm} values for the vocal tract based system and the proposed system based on excitation source are compared in Table 1.

Table 1. Speaker segmentation performance of the vocal tract and excitation source based systems ($C_{default} = 0.39$).

System based on	FAR	MDR	C_{seg}	C_{norm}
Vocal tract features	52%	64%	0.35	0.90
Excitation source features	37%	48%	0.27	0.69

6 Summary

In this paper, we have shown the effectiveness of subsegmental features for speaker change detection. Experiments with current approaches indicate that speaker segmentation methods based on statistical distribution of feature vectors do not perform satisfactorily when speaker turns are short (< 5 s). Excitation source features present in the LP residual of speech signal are useful for segmentation. The features can be extracted using AANN models. The results indicate that the subsegmental features from the excitation source signal perform better than the features representing the vocal tract. Combining evidences from multiple AANN models is still an issue and more exploration on this part may lead to improved performance.

References

1. Chen, S., Gopalakrishnan, P.: Speaker, environment and channel change detection and clustering via the Bayesian information criterion. In: Proceedings of DARPA Broadcast News Transcription and Understanding Workshop. (1998) 127–132
2. Delacourt, P., Wellekens, C.J.: DISTBIC: A speaker-based segmentation for audio data indexing. Speech Communication **32** (2000) 111–126
3. H. Gish, M. Siu and R. Rohlicek: Segregation of speakers for speech recognition and speaker identification. In: Proceedings of the International Conference on Acoustics Speech and Signal Processing. Volume 2. (1991) 873–876
4. Makhoul, J.: Linear prediction: A tutorial review. Proceedings of the IEEE **63** (1975) 561–580
5. Rabiner, L., Juang, B.H. In: Fundamentals of Speech Recognition. Prentice-Hall Inc., (Englewood Cliffs, New Jersey, USA)
6. B. Yegnanarayana and K. Sharat Reddy and S. P. Kishore: Source and system features for speaker recognition using aann models. In: Proceedings of the International Conference on Acoustics Speech and Signal Processing. Volume 1. (2001) 409–412
7. Yegnanarayana, B., et. al.: IIT Madras Speaker Recognition system. In: Proc. NIST Speaker Recognition Workshop, Baltimore, Maryland, USA (2003)