# USE OF VOCAL SOURCE FEATURES IN SPEAKER SEGMENTATION

*W.N. Chan, Tan Lee, Nengheng Zheng and Hua Ouyang*

Department of Electronic Engineering, the Chinese University of Hong Kong,
Shatin, New Territories, Hong Kong
Email: {wnchan, tanlee, nhzheng, houyang}@ee.cuhk.edu.hk

## ABSTRACT

This paper addresses the problem of speaker segmentation in telephone conversation. The segmentation is done in three steps: 1) preliminary segmentation to hypothesize speaker turning points; 2) clustering of segments; and 3) re-segmentation to determine speaker identity of each segment. It is found that vocal source related features are more speaker-discriminative than the conventional vocal tract related features for small amount of data. This motivates us to thoughtfully incorporate vocal source features into early stages of the speaker segmentation process, where decisions have to be made with limited data. Speaker segmentation experiments are carried out on 36 summed channel conversations in the NIST 2004 Speaker Recognition Evaluation. The proposed use of vocal source features leads to noticeable performance improvement.

## 1. INTRODUCTION

Speaker segmentation is a task of dividing an input speech signal into homogenous regions, each of which contains the speech of exactly one speaker. One of the applications is audio indexing and searching [1]. For example, the audio recording of a meeting or a conversation can be indexed automatically to facilitate the search and retrieval of the content spoken by a specific person. Speaker segmentation techniques are also very useful for automatic labeling and transcription of audio archives that involve multiple speakers [2,3,4]. In this application, the audio signal typically contains speech from different speakers under different acoustic conditions. With the knowledge of "who is speaking", acoustic models for speech recognition can be adapted to better match the conditions and the speakers. In the speech-to-text conversion process, information about speaker turns can be used to avoid linguistic discontinuity [4].

There are two basic problems to be addressed in automatic speaker segmentation. First, speaker turning points, i.e. the time instants when there are changes of speakers, need to be determined. Second, the speech segments separated by the turning points are associated with different speakers. In a typical application of speaker segmentation, speaker models can not be built in advance because the speaker identities are unknown and no speech data is available for training [1]. Therefore, a segment clustering process is required to aggregate similar segments, which are supposed to be from the same speaker.

Speaker turning point detection and segment clustering can be done sequentially in one pass [3,5] or iteratively in multiple passes [2,6,7]. Turning points are hypothesized based on local change of acoustic properties. The hypothesized turning points divide the speech signal into many segments. These segments are clustered into a certain number of speaker-homogenous groups. Statistical modeling techniques are commonly used for both turning point detection and segment clustering. The effectiveness of statistical modeling depends greatly on the amount of available training data. In the task of speaker segmentation, data is usually limited. In this paper, we consider two-speaker telephone conversations in which the duration of speech are short (1 to 3 seconds). This makes statistical modeling difficult.

In previous research on speaker segmentation, the most commonly used acoustic features are cepstral coefficients computed by filter-bank or linear prediction analysis [1,2,3,5,6,7]. These features describe mainly vocal tract related information and provide useful acoustic cues for phonetic classification i.e. speech recognition. In other words, they contain a great deal of content-specific information. For telephone conversation, each speaker speaks for a short period of time, typically. When determining the speaker turning points, a short window of speech is searched [1]. Within this window, the linguistic content is very limited. A statistical distribution computed from vocal tract features of such a short window tends to be biased by the specific linguistic content and thus be less effective in characterizing the speaker's voice. On the other hand, vocal source excitation signal carries useful speaker-specific information, e.g. pitch, types of glottal pulses, degree of breathy or creaky voice [8]. Compared with vocal tract features, vocal source features are less volatile to the variation of phonetic content. They are expected to be more appropriate for speaker discrimination when there is little speech data.

It is not a trivial task to extract useful vocal source features from acoustic signals. Linear prediction (LP) residual signal is considered as a useful manifestation of vocal source excitation, which can be computed efficiently [8]. Our previous work proposed to apply pitch-synchronous wavelet transform to LP residual signal. The resulted features, named Wavelet Octave Coefficients of Residues (WOCOR), are found to provide complementary discrimination power to the conventional vocal tract features [9].

In this paper, it is shown that WOCOR is more speaker-discriminative than the conventional Mel-frequency cepstral coefficients (MFCC) for short speech segments. We propose to use WOCOR as the primary acoustic features for speaker turning point detection. This leads to a significant improvement on the overall performance of speaker segmentation.

## 2. BASELINE SYSTEM

### 2.1. Speech Database

We use part of the speech data provided for the NIST 2004 Speaker Recognition Evaluation. The data is for the task of speaker detection in summed-channel conversations. Each conversation involves two speakers talking over telephones. It was created by sample-by-sample addition of the two sides of conversations [10]. The speech was sampled at *8* kHz and encoded by *8*-bit μ-law. The

duration of each conversation is approximately *5* minutes. No prior information about gender is provided.

There are about 1,200 conversations defined for training and testing in the NIST evaluation. In this research, 36 randomly selected conversations are used. The two speakers in a conversation could be both male, both female, or one male with one female. The language being spoken is English,Mandarin and mixture of them.

Each of the selected conversations was manually divided into speaker-homogeneous segments. The human annotator was allowed to visualize the waveform and listen to the audio signal back and forth before making decisions on turning point locations. The manual segmentation results are used as the reference for evaluation of the proposed speaker segmentation algorithms. A total of 1,857 speaker segments are marked in the 36 conversations. Excluding the silence and non-speech periods, the segment duration is mostly between 1 to 3 seconds.

### 2.2. Automatic Speaker Segmentation

Given one of the conversations, automatic speaker segmentation is performed in the following steps:

Step 1 – preliminary segmentation [1,3,5]

Step 2 – segment clustering [3,5]

Step 3 – re-segmentation [6,7]

These procedures are based on the methods previously proposed in [1,3,5,6,7]. But in our case, the number of speakers is known. The details of each step are described below.

*2.2.1. Preliminary segmentation*

Preliminary segmentation is to find a set of hypothesized speaker turning points. This is done with the DISTBIC technique proposed by Delacourt [1], which involves sequential use of spectral distance measurement and Bayesian information criterion (BIC).

When applying spectral distance measurement, we consider a 2-second window of speech in each measurement. The window is divided into two equal parts and each of them is represented by a single Gaussian distribution [1]. Let *f* and *g* denote the two distributions respectively. The Kullback-Leibler distance (KLD) is computed as follows,

$$KLD(f,g)=\frac{1}{2}tr\{(\Sigma_f^{-1}+\Sigma_g^{-1})(\mu_f-\mu_g)(\mu_f-\mu_g)^T+\Sigma_f\Sigma_g^{-1}+\Sigma_g\Sigma_f^{-1}-2I\} \quad (1)$$

As the window slides over the speech signal, the KLD is computed as a function of time. The window shift is 100 ms. The peak values on the time-varying KLD curve suggest the presence of the speaker turning points.

Subsequently, the turning points are refined using ΔBIC value [1]. Let $X=\{x_1, \ldots, x_N\}$ be the feature vector sequence from N successive frames of speech. In this study, N is set to 200 (2 seconds). Consider the two sub-sequences $X_1=\{x_1, \ldots, x_i\}$ and $X_2=\{x_{i+1}, \ldots, x_N\}$, where $1 < i < N$. The following two hypotheses are defined [3]:

$H_0$ — X is generated by a single Gaussian distribution detnoted by $N(\mu,\Sigma)$, where μ and Σ are the mean and full covariance matrix respectively

$H_1$ — $X_1$ and $X_2$ are generated by two distinct Gaussian distributions $N(\mu_1,\Sigma_1)$ and $N(\mu_2,\Sigma_2)$ respectively

ΔBIC value is given by the likelihood ratio of $H_0$ and $H_1$, minus a penalty [3], i.e.

$$\Delta BIC(i)=N\log|\Sigma|-N_1\log|\Sigma_1|-N_2\log|\Sigma_2|-\lambda P \quad (2)$$

The penalty term is used to balance the model complexity difference. It is defined as,

$$P=\frac{1}{2}(d+\frac{1}{2}d(d+1))\log N \quad (3)$$

where d is the dimension of feature vectors $x_i$. The turning point is detected by

$$t=\arg\max_i \Delta BIC(i) \quad (4)$$

If $\Delta BIC(i)\le 0$ for all *i*, no turning point is assumed for X.

λ controls the sensitivity of turning point detection [1,3]. In our case, a low miss rate is desired without seriously compromising the false alarm rate, because missed turning points are not recoverable afterwards.

*2.2.2. Segment clustering*

The speaker turning points found as described above divide the conversation into many segments, each of which is assumed to be speaker-homogeneous. These segments are merged using hierarchical clustering technique [3,5]. This is a bottom-up method. The distance between each possible pair of clusters is calculated as in Eq.(2). The closest pairs are merged to form a new cluster. Given that we are tackling a two-speaker task, the clustering process stops when there are only two clusters left.

*2.2.3. Re-segmentation*

A Gaussian mixture model with 32 mixture components is trained to represent each of the two clusters, which correspond to the two speakers. Viterbi re-segmentation of the conversation is performed using these models. It determines the most probable path that toggles between the two speaker states [6,7]. There is a constraint of minimum segment length of 0.5 seconds. Unlike what is suggested in [6,7], there is no further iteration of clustering and re-segmentation. Only one pass of re-segmentation is performed to produce the final result of speaker segmentation.

*2.2.4. Performance measurement*

Turning point detection

A false alarm (FA) of turning point detection occurs when a detected turning point is not a true one. A missed detection (MD) occurs when a true turning point can not be detected. The false alarm rate and missed detection rate are defined as,

$$FAR=\frac{\sum FA_i}{\sum FA_i+\sum turn_i}\times 100\%,\quad MDR=\frac{\sum MD_i}{\sum turn_i}\times 100\%$$

where $\sum FA_i$ and $\sum MD_i$ are the total number of FA and MD respectively, $\sum turn_i$ is the total number of true turning points given by the reference manual segmentation. The turning point detection rates can be evaluated either on the preliminary segmentation or on the re-segmentation result.

Speaker coverage

The segmentation performance can be assessed in terms of speaker coverage, i.e.

$$MRCov=\frac{\sum_\theta \text{duration of missed portion for reference segment } \theta}{\sum_\theta \text{duration of reference segment } \theta}\times 100\%$$

$$FACov=\frac{\sum_\rho \text{duration of false portion for detected segment } \rho}{\sum_\rho \text{duration of detected segment } \rho}\times 100\%$$

*2.2.5. Baseline results*

The acoustic features used in the baseline system include *12* MFCCs and the log energy, which are computed with a Hamming window of *20* ms long with *10* ms frame shift. The baseline performance of speaker segmentation is shown as in Table 1. For preliminary segmentation, FAR increases when MDR decreases. The thresholds for KLD peak detection and the parameter $\lambda$ in Eq.(2) can be adjusted to reach a specific operating point. In our application, we wish to have the MDR as low as possible because the missed segment can not be recovered. In the following experiments, these parameters are empirically determined by trial tests with two selected conversations. The resulted parameter values are applied to all conversations.

The GMM based re-segmentation is very effective in reducing both the false alarm rate and missed detection rate of turning point detection. It is found that about 45% of the segmentation errors are attributed to segments of 1 - 5 seconds in length. To deal with such short segments, the MFCC features may not be appropriate because they may be biased by specific linguistic content in the segment. In the following section, we show that vocal source excitation features are more discriminative for short speech segments.

**Table 1.** Baseline performance of speaker segmentation

| | | | |
|---|---|---|---|
| Turning point detection accuracy | Preliminary segmentation | FAR | 70.7% |
| | | MDR | 47.0% |
| | Re-segmentation | FAR | 31.6% |
| | | MDR | 25.9% |
| **Speaker coverage (after re-segmentation)** | | **FACov** | **12.3%** |
| | | **MRCov** | **9.00%** |

## 3. VOCAL SOURCE FEATURES: WOCOR

It is believed that vocal source excitation features contain important speaker-specific information. The LP residual signal has been regarded as an effective carrier of vocal source information [8,9]. In our previous work [9], it was proposed to apply pitch-synchronous wavelet transform to the LP residual signal to derive a set of vocal source features, named WOCOR, for speaker recognition. Only voiced frames are used to derive WOCOR. It was shown that WOCOR provides additional speaker discriminative power to the commonly used Linear Predictive Cepstral Coefficients (LPCC) [9].

*3.1 Computation of WOCOR*

The residual signal is generated by linear predictive inverse filtering and pitch epochs are identified for synchronization. Then, 3-level discrete wavelet transform (DWT) is applied to every two successive pitch cycles of the residual signal, resulting in three groups of detail coefficients and one group of approximation coefficients. Each group of coefficients represents a specific spectral component. Each group is further divided into four sub-groups and the overall energy of each sub-group is computed as a feature component. The entire WOCOR feature vector contains *16* components. By multi-level DWT, the pitch-related low-frequency properties and high-frequency information associated with pitch epochs are captured with different resolutions of time-frequency analysis. Dividing each group into sub-groups enables the characterization of temporal variation of the spectral components within a pitch period and that over consecutive periods. Therefore, WOCOR is capable of capturing the spectro-temporal characteristics of the LP residual signal [9].

*3.2. Discrimination power of WOCOR against MFCC*

The discrimination power of WOCOR is analyzed in comparison with the conventional MFCC. We are interested to know whether WOCOR would be a more effective feature for speaker discrimination when the number of speech samples is small. This simulates the difficulty of working with insufficient data for speaker turning point detection.

Let $S_{ij}$ denote the j[th] segment of speaker i (i = 1 or 2) in a given conversation. The reference manual segmentation is used here. We attempt to build a speaker model based on M randomly selected feature vectors from this segment. Let these selected feature vectors be denoted by $F_{ij}$ . The features are either MFCC or WOCOR. By varying M, the amount of data for statistical modeling can be controlled and tested.

Without loss of generality, we first consider speaker 1 to be the target speaker and speaker 2 be the imposter, and perform a speaker verification test. For the segment $S_{1j}$ , the target model for speaker 1, denoted by $G_1$ , is trained by $F_{1j}$ and the anti-model, denoted by $G_2$ , is trained by $F_{2k}$ , which is extracted from a randomly selected segment $S_{2k}$ of speaker 2.

All segments from speaker 1 except $S_{1j}$ are used for the test. For each feature vector, the likelihoods produced by $G_1$ and $G_2$ are computed. The overall likelihood of a segment is the average likelihood of all feature vectors in the segment.

Let $p(F_{1p} \mid G_1)$ and $p(F_{1p} \mid G_2)$ be the segment-level likelihoods for $S_{1p}$ where p ≠ j. The likelihood ratio is computed as,

$$\mathrm{LR}_{1\mathrm{p}} = \log p(F_{1p} \mid G_1) - \log p(F_{1p} \mid G_2) \quad (5)$$

If $LR_{1p} > 0$, we consider that this is a correct decision of speaker verification.

The same testing procedures are repeated for all segments of the two speakers. Suppose that there are L segments from a speaker, a total of L×(L-1)/2 tests can be performed for the speaker. A large number of tests are carried out for the 36 conversations and a statistical error rate is obtained.

The performance of MFCC and WOCOR are evaluated with different values of M, i.e. the number of feature samples used to represent a segment. If there are less than M feature samples in a segment, the segment is not used for testing.

Figure 1 shows the performance of WOCOR and MFCC, with M varying from 50 to 200, which are equivalent to speech duration of 0.5 to 2 seconds. WOCOR performs better than MFCC when the number of feature samples is small, i.e. M < 110 (or 1.1 seconds). With the amount of data increasing, MFCC begins to overtake WOCOR. When M is greater than 170, MFCC becomes obviously better than WOCOR.

It is noted that, for some of the male-female conversations, WOCOR works better than MFCC for segment length as long as 2 seconds (M=200). However, for female-female conversations, the effectiveness of WOCOR is not noticeable unless M becomes very small, i.e. M < 90. For most male-male conversations, the MFCC overtakes WOCOR around 90 and the error rate is high.

WOCOR contains pitch-related information and therefore is good at distinguishing speakers of opposite genders. This may explain the superiority of WOCOR in male-female conversation. For speakers of the same gender, pitch information becomes less discriminative. For female speakers, pitch periods are relatively short and pitch extraction becomes less accurate. This affects the pitch synchronization and makes WOCOR less reliable.

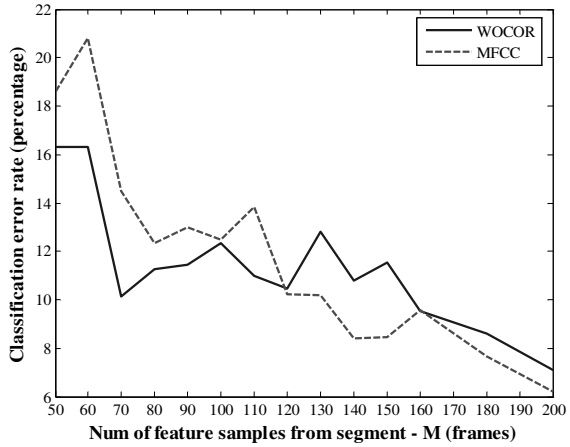


**Fig.1.** Speaker classification performance attained by WOCOR and MFCC with different number of feature samples

## 4. USE OF WOCOR IN SPEAKER SEGMENTATION

WOCOR is more effective than MFCC for speaker discrimination with relatively small amount of data. MFCC outperforms WOCOR if the feature samples are abundant. For speaker turning point detection, we replace MFCC by WOCOR in the computation of both KLD and BIC. For segment clustering, if the total amount of data in a cluster is equivalent to 2 seconds of speech or shorter, a linear score fusion of MFCC and WOCOR is adopted for the BIC clustering, i.e.

$$\Delta BIC_{fused} = w_{MFCC}(\Delta BIC_{MFCC}) + w_{WOCOR}(\Delta BIC_{WOCOR}) \quad (6)$$

where $w_{MFCC}$=*0.3* and $w_{WOCOR}$=*0.7* are determined empirically. If the cluster contains more than 2 seconds of speech, the clustering decision depends solely on MFCC.

The experimental results are shown as in Table 2. Similar to the baseline experiments, the thresholds for KLD peak detection and the parameter $\lambda$ in Eq.(2) are empirically determined by trial tests with two selected conversations. The same parameter values are applied to all conversations.

**Table 2.** Speaker segmentation performance with WOCOR used in speaker turning point detection. ( ) gives the baseline performance

| | | | |
|---|---|---|---|
| Turning point detection accuracy | Preliminary segmentation | FAR | 73.9% (70.7%) |
| | | MDR | 21.1% (47.0%) |
| | Re-segmentation | FAR | 27.3% (31.6%) |
| | | MDR | 20.7% (25.9%) |
| **Speaker coverage (after re-segmentation)** | | **FACov** | **9.94% (12.3%)** |
| | | **MRCov** | **7.33% (9.00%)** |

For preliminary segmentation (turning point detection), the use of WOCOR leads to a 25.9% absolute reduction of the missed detection rate (MDR) while the false alarm rate (FAR) increases slightly by 3.2%. This verifies our earlier observation that WOCOR is more effective than MFCC for discriminating speakers with relatively small amount of data.

After re-segmentation, FACov improves from the baseline performance of 12.3% to 9.9%; MRCov also drops from 9.0% to 7.3%. Although the preliminary segmentation performance is much improved by the proposed use of WOCOR, the subsequent segment clustering and GMM based re-segmentation rely also on MFCC. Moreover, the amount of data used in segment clustering and re-segmentation is relatively large. The advantage of WOCOR becomes less noticeable in this case.

## 5. CONCLUSIONS

The major finding of this paper is that vocal source related features, namely WOCOR, are more speaker-discriminative than the conventional MFCC features for small amount of data. This has motivated us to thoughtfully incorporate vocal source features into early stages of the speaker segmentation process, where decisions have to be made with limited data. The effectiveness of our proposed way of using vocal source features is confirmed favorably by a noticeable performance improvement on speaker segmentation.

## ACKNOWLEDGEMENT

This research was partially supported by a Central Allocation Grant (Ref. CUHK1/02C) from the Hong Kong Research Grants Council.

## REFERENCES

[1] P. DelaCourt and C.J. Wellekens, "DISTBIC: A speaker-based segmentation for audio data indexing," *Speech Communications*, Vol.32, pp.111-126, 2000.

[2] J. Ajmera and C. Wooters, "A robust speaker clustering algorithm", *Proceedings of 2003 IEEE Automatic Speech Recognition and Understanding*.

[3] S.S. Chen and P.S. Gopalakrishnan, "Speaker environment and channel change detection and clustering via the Bayesian information criterion," *Technical Report, IBM T.J. Watson Research Center*, 1998.

[4] J.-L. Gauvain, L. Lamel, G. Adda and M. Jardino, "The LIMSI 1998 Hub-4E transcription system," *Proceedings of DARPA Broadcast News Workshop*, 1999.

[5] S.E. Tranter and D.A. Reynolds, "Speaker diarization for broadcast news," *Proceedings of Odyssey Speaker and Language Recognition Workshop*, 2004.

[6] C. Barras, X. Zhu, S. Meignier and J.-L. Gauvain, "Improving speaker diarization," *Proceedings of Fall 2004 Rich Transcription Workshop (RT-04F)*, 2004.

[7] J.-L. Gauvain, L. Lamel and G. Adda, "Audio partitioning and transcription for broadcast data indexation," *Multimedia Tools and Applications*, Vol.14, pp.187-200, 2001.

[8] M.D. Plumpe, T.F. Quatieri and D.A. Reynolds, "Modeling of the glottal flow derivative waveform with application to speaker identification," *IEEE Trans. Speech Audio Processing*, Vol.7, No.5, pp.569-585, 1999.

[9] Nengheng Zheng, P.C. Ching and Tan Lee, "Time-frequency analysis of vocal source signal for speaker recognition," *Proceedings of ICSLP* 2004, pp.2333-2336.

[10] The NIST Year 2004 Speaker Recognition Evaluation Plan, http://www.nist.gov/speech/tests/spk/2004/SRE-04_evalplan-v1a.pdf