TEMPORAL DECOMPOSITION: A PROMISING APPROACH TO VQ-BASED SPEAKER IDENTIFICATION

Phu Chien Nguyen, Masato Akagi, and Tu Bao Ho

Japan Advanced Institute of Science and Technology
1-1 Asahidai, Tatsunokuchi, Ishikawa 923-1292, Japan
{chien, akagi, bao}@jaist.ac.jp

ABSTRACT

In this paper, a new set of features is proposed that has been found to improve the performance of automatic speaker identification systems. The new set of features is referred to as "event targets". The new features have been derived from line spectral frequency (LSF) parameters using the so-called "temporal decomposition" (TD) technique. The number of feature vectors required for both training and testing phases has been reduced by one-fifth compared to that of the traditional mel-frequency cepstrum coefficients (MFCC) features, while the identification results obtained are comparable or even better. Also, this work introduces one more application of TD (speaker recognition) in addition to speech coding, speech segmentation, and speech recognition. It shows that the event targets in TD can convey information about the identity of a speaker.

1. INTRODUCTION

Speaker recognition is the process of automatically recognizing the person speaking based on individual information included in speech waves. There are two types of tasks within speaker recognition: identification and verification. The objective of a speaker identification (ID) system is to determine the identity of an individual from a sample of his or her voice. Speaker ID can be further subdivided into two categories: closed set or open set. A closed-set speaker ID system identifies the speaker as one of those enrolled, even if he or she is not actually enrolled in the system. On the other hand, an open-set speaker ID system should be able to determine whether a speaker is enrolled or not, if enrolled, determine his or her identity [1].

Another distinguishing aspect of speaker recognition systems is that they can be either text-dependent or text-independent. In the text-dependent case, the input sentence or phrase is fixed for each speaker, whereas in the text-independent case, there is no restriction on the sentence or phrase to be spoken. Speaker ID consists of two stages, namely, feature extraction and classification as shown in Fig. 1. This paper focuses on the feature extraction aspect of the problem of text-independent closed-set speaker ID.

Feature extraction is the process of deriving a compact set of parameters that are characteristics of a given speaker. Ideally, these parameters should efficiently preserve all the information relevant to the speaker's identity while eliminating any irrelevant information. That is, they should minimize the intra-speaker variance and at the same time maximize the inter-speaker variances [1]. The majority of speaker recognition systems use some types of short-time spectral analysis. The most effective and widely used spectral analysis techniques for speaker recognition are linear prediction (LP) analysis [1, 2] and filter bank analysis [9, 12]. This paper focuses on LP-derived features, namely, "event targets" that are extracted from the line spectral frequency (LSF) parameters using the so-called "temporal decomposition" (TD) technique [3, 6, 8, 11].

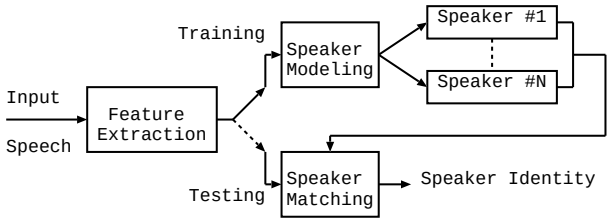

Fig. 1. Block diagram of speaker identification systems.

The state-of-the-art in classification techniques used in speaker recognition include Dynamic Time Warping (DTW), Hidden Markov Modeling (HMM), Gaussian Mixture Modeling (GMM), and Vector Quantization (VQ) [7]. In this work, the VQ-based speaker identification is used, due to ease of implementation and high accuracy. It is well-known that the VQ approach has demonstrated good performance on limited vocabulary tasks. However, this method is somewhat impractical when the number of training and/or testing vectors is large, since the memory and amount of computation required become prohibitively high. Alternatively, event targets as a new set of features for speaker recognition can help to alleviate this problem.

The rest of the paper is organized as follows. In Section 2, we briefly review the baseline VQ-based speaker ID. Next, the process of event target extraction is described in Section 3 and experimental results are reported in Section 4. Finally, conclusions are drawn in Section 5.

2. VQ-BASED SPEAKER IDENTIFICATION

Vector quantization (VQ) is a process of mapping vectors from a large vector space to a finite number of regions in that space. Each region is called a cluster and can be rep-

resented by its center called a codeword. The collection of all codewords is called a codebook.

VQ is used in both training and matching phases of a VQ-based speaker ID system. In the training phase, a speaker-specific VQ codebook is generated for each known speaker by clustering his or her training acoustic vectors. The distance from a vector to the closest codeword is called a VQ-distortion. In the matching phase, an input utterance of an unknown speaker is vector quantized using each trained codebook and the total VQ distortion is computed. The speaker corresponding to the VQ codebook with smallest total distortion is identified [7, 13].

3. EXTRACTION OF EVENT TARGETS

In articulatory phonetics, speech production is considered as a sequence of overlapping articulatory gestures, each of which may be thought of as a movement towards and away from an ideal, but often not reached, articulatory target. The sound produced by such an articulatory movement corresponds to a phoneme or a sub-phoneme in speech. In other words, each gesture produces an acoustic event that should approximate a phonetic target. Adjacent gestures overlap one another resulting in the characteristic transitions between phonemes that can be observed in almost any parametric representation of the acoustic speech signal. Due to co-articulation and reduction in fluent speech, a target may not be reached before articulation towards the next phonetic target begins. It has long been a difficult task to determine such targets and their temporal evolutionary patterns from the acoustic signal alone.

The temporal decomposition (TD) method for analyzing speech achieves the objective of decomposing speech into targets and their temporal evolutionary patterns, without any recourse to any explicit phonetic knowledge [3]. This model of speech takes into account the above articulatory considerations and results in a description of speech in terms of event targets describing the ideal articulatory configurations of the successive acoustic events in speech, and event functions describing their temporal evolutionary patterns. Therefore, it tries to achieve an optimal transformation from the multidimensional spectral parameter space to the phonetic space which can be considered for many applications to be a powerful speech analysis technique.

Suppose that a given utterance has been produced by a sequence of K movements aimed at realizing K acoustic targets. Let us denote the speech parameters corresponding to the kth target by $\mathbf{a}(k)$, and the temporal evolution of this event by a function, $\phi_k(n)$. The frame number n varies between 1 and N. In temporal decomposition of speech, the observed speech parameters, $\mathbf{y}(n)$, are approximated by $\hat{\mathbf{y}}(n)$, a linear combination of event targets as follows.

$$\hat{\mathbf{y}}(n) = \sum_{k=1}^{K} \mathbf{a}_k \phi_k(n), \qquad 1 \le n \le N \tag{1}$$

In matrix notation the Equation (1) can be written as

$$\hat{\mathbf{Y}} = \mathbf{A}\mathbf{\Phi} \qquad \hat{\mathbf{Y}} \in R^{P\times N}, \mathbf{A} \in R^{P\times K}, \mathbf{\Phi} \in R^{K\times N}$$

where P is the dimension of the spectral parameters. In Equation (1), both the event targets and event functions

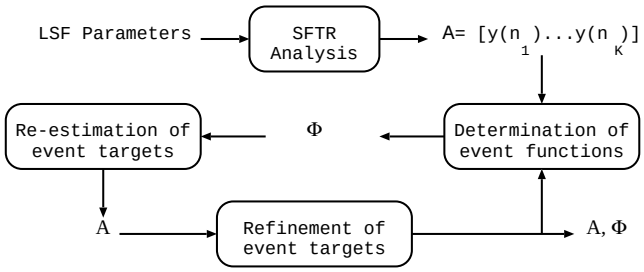

Fig. 2. Block diagram of the MRTD algorithm.

are unknown and the temporal decomposition analysis involves the determination of them once the speech parameter sequence of an utterance is given.

The restricted second order TD model was utilized in [6, 8, 11], where only two adjacent event functions can overlap and all event functions at any time sum up to one. The argument for imposing this constraint on the event functions can be found in [6, 11]. Equation (1) is rewritten as

$$\hat{\mathbf{y}}(n) = \mathbf{a}_k \phi_k(n) + \mathbf{a}_{k+1}(1 - \phi_k(n)), \quad n_k \le n < n_{k+1} \tag{2}$$

where n_k and n_{k+1} are the central positions of event k and event $k+1$, respectively.

In order to apply TD to decomposing line spectral frequency (LSF) parameters, the stability of the corresponding linear predictive coding (LPC) synthesis filter after spectral transformation performed by TD must be ensured. The restricted temporal decomposition (RTD) method [8] intends to make LSF parameters possible for TD by enforcing the LSF ordering property on the event targets. However, RTD has not completely solved this problem as indicated in [11]. Moreover, some event functions derived from RTD are ill-shaped, i.e. they have more than one peak, which is undesirable from speech coding point of view. Thus, the modified RTD (MRTD) method [11] has been proposed to overcome the drawbacks imposed on the RTD method. The block diagram of MRTD is shown in Fig. 2 and the whole algorithm is summarized as follows. For a detailed mathematical treatment, the reader is referred to [8, 11].

First, the initial approximation of event targets is based on a maximum spectral stability criterion. It is assumed that each acoustic event that exists in speech gives rise to a spectrally stable point in its neighborhood. Therefore, the locations of the spectrally stable points and the corresponding spectral parameter vectors can be used as a good approximation to the event locations and event targets, respectively. Here, event localization is done via the local minimal points of a spectral transition measure called spectral feature transition rate (SFTR).

In the result, when once the locations of events n_k, where $k = 1, ..., K$, are known and the corresponding event targets are initialized with the samples of the LSF vector trajectory $\mathbf{y}(n_k)$, we can calculate proper event functions and event targets iteratively in the least mean square sense. However, since the event targets are calculated using the formula $\mathbf{A} = \mathbf{Y}\mathbf{\Phi}^T(\mathbf{\Phi}\mathbf{\Phi}^T)^{-1}$ which does not consider the LSF ordering property for them, the estimated event targets

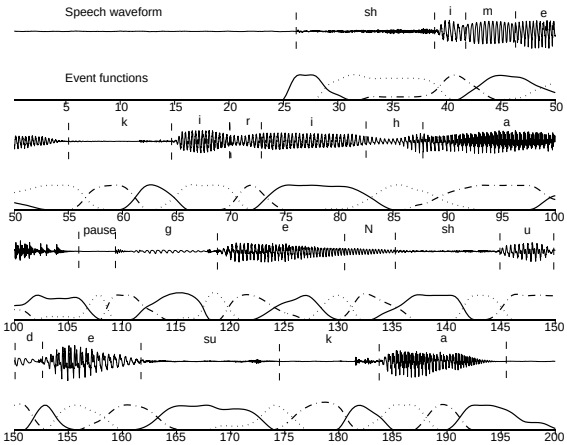

Fig. 3. Plot of the event functions obtained from MRTD for the Female/Japanese speech utterance "*shimekiri ha geN-shu desu ka*". The speech waveform is also shown together with the phonetic transcription for reference. The numerals indicate the frame numbers.

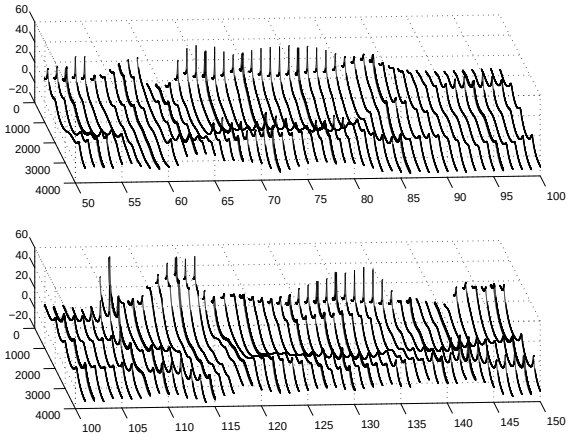

Fig. 4. Event targets obtained from MRTD for a segment of the Female/Japanese speech utterance "*shimekiri ha geN-shu desu ka*" (from frame number 51 to 150). Dark solid lines show the log power spectra of event targets. The log power spectra of the original LSF vectors are also provided.

may not be interpreted as LSF vectors. Invalid LSF event targets estimated from a LSF vector trajectory cause two serious problem. Firstly, the event targets do not have their own spectra as valid LSF vectors do. It follows that those event targets are regarded as the numerical results, but not as the idealized targets. They also prohibit us from matching the determined events with meaningful phonetic units. Secondly, it is impossible to utilize the advantages of LSF parameters for quantization. The invalid LSF event targets lower the intra/inter-correlations and do not guarantee the stability of the reconstructed LSF vectors. Therefore, a refinement procedure is applied to the estimated event targets to ensure the LSF ordering property for them with a negligible increase in reconstruction error.

Fig. 3 shows the plot of event functions obtained from the MRTD method for an example of a Female/Japanese speech utterance. The associated event targets obtained from MRTD analysis for a segment of the same speech utterance are shown in Fig. 4.

The concept of temporal decomposition of speech has attracted many researchers in recent years, specially in application areas such as speech coding, recognition and segmentation. The fact that TD decomposes the speech parameters into two elementary components, which occur at a lower rate than the original speech parameters, gives a means of coding speech efficiently at a lower bit rate [3, 8, 11]. The strong relationship between the TD representation of speech and the speech production mechanism has provided the necessary motivation to investigate its application in speech recognition [4, 5]. Its usefulness in speech segmentation has also been investigated [6].

The application of TD to VQ-based speaker ID is motivated by the fact that TD is promising as a means of segmenting speech into a sequence of overlapping events closely related to phonetic structure of the speech signals [5]. On the other hand, the VQ-based speaker ID can be regarded as a method that use phoneme-class-dependent speaker characteristics in short-term spectral features through implicit phoneme-class recognition. In other words, phoneme-classes and speakers are simultaneously recognized in this method [7]. Therefore, the event targets extracted from spectral parameters using TD can be considered as a new set of features for VQ-based speaker ID.

Table 1. Summary of the speaker set.

# Speakers	49 (31 M+18 F)
Avg. duration of training utterance	24.5 sec/speaker
Avg. duration of testing utterance	3.1 sec/sentence

4. EXPERIMENTAL RESULTS

4.1. Database

In the experiments, we used a speaker set of 49 speakers collected from the New England dialect of TIMIT speech corpus. The ratio of male and female speakers is not equal in the set. For each speaker, there are ten sentences. The training set is generated using the eight files with "sx" and "si" prefixes, whereas the two files with "sa" prefix are individually used for testing. Summary of the speaker set is given in Table 1.

4.2. Preprocessing and Feature Extraction

Prior to any analysis, the speech files were downsampled from 16 to 8 kHz. High emphasis filtering with $H(z) = 1 - 0.95z^{-1}$ was then performed.

To derive event targets used for VQ-based speaker ID, 10^{th} order LSF parameters were calculated first, using a LPC analysis window of 30 ms at 10 ms frame intervals. In the following, the LSF parameters obtained were TD analyzed using the MRTD method. The event rate, i.e. the number of events per second, was set as about 20 events per

Table 2. Total number of feature vectors used in the experiments.

Phase	# MFCC Vectors	# Event Targets
Training	118861	23511
Testing	30231	5929

Table 3. Identification success rates for different codebook sizes and feature sets. Note that LSF features were calculated at the event locations.

Codebook Size	MFCC	Event Targets	LSFs
16	89.80%	93.88 %	85.71 %
32	95.92%	95.92 %	93.88 %
64	94.90%	96.94 %	95.92 %
128	95.92%	96.94 %	95.92 %
256	95.92%	97.96 %	93.88 %

second, resulting in the number of event targets reduced by one-fifth compared to that of the original LSF vectors.

For comparison, conventional mel-frequency cepstrum coefficients (MFCC) were computed using the 12^{th} short-term mel-cepstrum analysis, also with a 30 ms Hamming window shifted by 10 ms, producing 100 feature vectors per second. The 12 lowest coefficients (excluding the 0^{th} coefficient, which corresponds to the total energy of the frame) were used as alternative features.

Table 2 gives the summary of feature vectors used in the experiments. It can be seen from the table that the number of event targets has significantly reduced compared to that of MFCC vectors in both training and testing phases.

4.3. Results

A separate classifier was used for each feature set. The distance measure here is the Euclidean distance. The codebooks for each speaker were designed using the LBG algorithm [10]. Speaker ID results for different codebook sizes and the two feature sets are given in Table 3. The performance of VQ-based speaker ID on the initiated event targets that consist of the original LSF vectors at event locations is also shown together for reference.

For all codebook sizes listed in the table, the event target features almost show better performance than the other features. This is mainly attributed to the fact that TD can be considered as an effective method of decorrelating the inherent inter-frame correlation present in any frame-based parametric representation of speech. In addition, results also show that the iterative refinement of event targets has positively affected their speaker-specific information.

5. CONCLUSIONS

The event targets derived from LSF parameters using the temporal decomposition technique were found to be effective when applied in a VQ-based speaker identification system. Their performance is found to be superior to that of the popular MFCC features in the case of testing on clean speech. The number of feature vectors required for both training and testing phases has been reduced by one-fifth compared to that of the MFCC features, while the identification results obtained are comparable or even better. More interestingly, it is shown that event targets can convey information about the identity of a speaker. We plan to make future experiments in more demanding environments such as testing on noisy speech, speech at different speaking rates, and cross-language evaluation.

6. REFERENCES

[1] K.T. Assaleh and R.J. Mammone, "New LP-Derived Features for Speaker Identification," *IEEE Trans. Speech and Audio Processing*, Vol. 2, No. 4, pp. 630-638, 1994.

[2] B.S. Atal, "Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification," *J. Acoust. Soc. Amer.*, Vol. 55, No. 6, pp. 1304-1312, 1974.

[3] B.S. Atal, "Efficient coding of LPC parameters by temporal decomposition," *Proc. ICASSP'83*, pp. 81-84, 1983.

[4] F. Bimbot, G. Chollet, P. Deleglise, and C. Montacie, "Temporal decomposition and acoustic-phonetic decoding of speech," *Proc. ICASSP'88*, pp. 445-448, 1988.

[5] A. M. L. Van Dijk-Kappers and S. M. Marcus, "Temporal decomposition of speech," *Speech Communication*, Vol. 8, No. 2, pp. 125-135, 1989.

[6] P.J. Dix and G. Bloothooft, "A breakpoint analysis procedure based on temporal decomposition," *IEEE Trans. Speech and Audio Processing*, Vol. 2, No. 1, pp. 9-17, 1994.

[7] S. Furui, "Recent advances in speaker recognition," *Pattern Recognition Letters*, Vol. 18, No. 9, pp. 859-872, 1997.

[8] S.J. Kim and Y.H. Oh, "Efficient quantization method for LSF parameters based on restricted temporal decomposition," *Electronics Letters*, Vol. 35, No. 12, pp. 962-964, 1999.

[9] T. Kinnunen, "Designing a speaker-discriminative adaptive filter bank for speaker recognition," *Proc. ICSLP'02*, pp. 2325-2328, 2002.

[10] Y. Linde, A. Buzo, and R.M. Gray, "An algorithm for vector quantiser design," *IEEE Trans. Communication*, Vol. 28, pp. 84-95, 1980.

[11] P.C. Nguyen and M. Akagi, "Improvement of the restricted temporal decomposition method for line spectral frequency parameters," *Proc. ICASSP'02*, pp. 265-268, 2002.

[12] D.A. Reynolds and R.C. Rose, "Robust text-independent speaker identification using Gaussian mixture speaker models," *IEEE Trans. Speech and Audio Processing*, Vol. 3, No. 1, pp. 72-83, 1995.

[13] F.K. Soong, A.E. Rosenberg, B.-H. Juang, and L.R. Rabiner, "A vector quantization approach to speaker recognition," *AT&T Technical Journal*, Vol. 66, No. 2, pp. 14-26, 1987.