ANALYSIS AND PROPOSING A ROBUST SOURCE FEATURES FOR AUTOMATIC TEXT-INDEPENDENT SPEAKER INDEXING SYSTEM

V.Subba Ramaiah[1] and R.Rajeswara Rao[2]

[1,2]Department of Computer Science & Engineering, MGIT, Hyderabad, AP, India
[1]subbubdl@gmail.com, [2]raob4u@yahoo.com

ABSTRACT

Speaker indexing (tracking) is the task of recognizing the multiple speakers from the given speech signal. Speaker indexing is a pattern recognition task. Every pattern recognition task is classified into three phases namely, feature extraction, training and testing phases. In this paper, we analyse and study about the major feature extraction techniques that exist and propose robust source features for text-independent speaker indexing system. A comparative study is carried out with each existing feature extraction techniques available with source features for speaker indexing task. Finally, we propose, source features are better than the other feature extraction techniques.

KEYWORDS

Speaker Indexing, Speaker Recognition, Source Feature, and Feature Extraction

1. INTRODUCTION

Speaker recognition is the process of automatically recognizing who is speaking on the basis of individual information included in speech waves. Speaker recognition can be classified into two tasks, namely speaker identification and speaker verification [1]. Speaker identification is the process of recognizing the speaker from the given spoken utterance from the registered set of 'N' Speakers. Speaker verification is the process of accepting or rejecting the identity claim of the speaker. Speaker recognizing technology is used in biometric approaches. This technology is used in many commercial applications such as banking by telephone, voice mail, and other security related applications. Speaker recognition is one of the centre field of research.

In existing systems, unknown speaker is recognized from the given speech signal. In many real-time conversations and news broad casting, the speech is continuous, beginning and end of speech segment of a speaker is unknown. Therefore, we need to index speech signals based on the speaker utterance. This process is called speaker segmentation or speaker indexing system. Speaker tracking is also essential in many applications, such as conference and meeting indexing [2], audio/video retrieval or browsing [3, 4], speaker adaptation for speech recognition [5,23], and video content analysis.

Traditionally, the speaker recognition task supposes that training and testing are composed of mono-speaker records. Then, to handle this kind of multi-speaker recordings, some extensions of the speaker recognition task are needed, such as:

DOI : 10.5121/cseij.2012.2401

- The N-speaker detection which is similar to speaker verification. It consists in determining whether a set of target speakers are speaking in a conversation.
- Speaker tracking is the task of recognizing the multiple speakers from the given speech signal.
- Speaker segmentation that is close to speaker tracking but there is no information about the identity and number of speakers present. The objective of this task is to determine the number of speakers in the given speech signal. This problem corresponds to a blind classification of the data, and the result is a partition in which every class is composed of segments of one speaker.

In this paper, we focus on the problem of speaker segmentation, detection and tracking in multi-speaker audio recordings using speaker biometrics. With the work presented here, our aim is to explore a new set of robust source features for speaker segmentation and speaker diarization system.

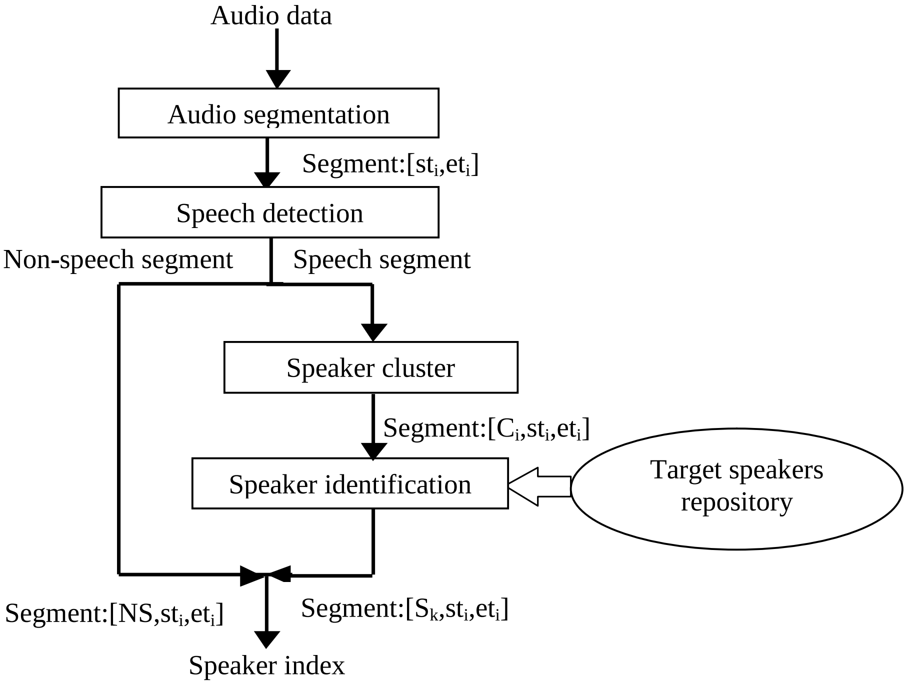

Figure 1. Block diagram of a typical speaker-indexing system

In Figure 1, the baseline speaker-indexing system architecture is depicted. First, the given speech signal is segmented and time-stamps are given to locations, where the changes are detected between the speakers. As shown in Figure 1, speech signals are segmented by labelling the starting and ending time of each segment (segment: [st_i, et_i]). The obtained speech segments are divided into speech and non-speech data. This is done in a speech segmentation module. In Figure 1, non-speech segments are marked as [NS, st_i, et_i], further non-speech segments are discarded. Then the clustering is done on speech signals. The objective of clustering is to classify the speakers. As shown in Figure 1, the speech segments [C_i, st_i, et_i] are clustered. In the next phase, speakers are identified using speaker identification module.

2. Characteristic of Robust Features for Speaker Tracking System

A set of desirable characteristics of feature vectors for speaker recognition are as given below [6]:

- Efficient in representing the speaker specific information
- Easy to measure
- Stable over time
- Occur frequently in speech
- Change little from one environment to another
- Not susceptible to mimicry
- High inter-speaker variations
- Low intra-speaker variations

For evaluating the features for their performance, a good measure of effectiveness is the F -ratio. It is defined as given below [7] [8] [9].

$$F = \frac{\mathrm{int}\,er - spea\,\mathrm{ker\ var}\,iation}{\mathrm{int}\,ra - spea\,\mathrm{ker\ var}\,iation} \qquad (1)$$

F -ratio should be high so as to have high discrimination efficiency for a given feature vector. Sometimes, by weighting the feature parameters according to their effectiveness, the performance of the system can be improved [10] [11] [12]. After extracting a set of effective feature vectors, speaker characteristics are captured by estimating the distribution of these feature vectors in the feature space. Methods used for capturing the distribution are explained in the next section.

3. Analysis of other Existing Feature Techniques

In this section, we review some of the approaches similar to our approach for text independent speaker recognition. Feature extraction [13] is the process of extracting the relevant information from the speech signal. The pitch and LPC-residual with the LPC-cepstrum are integrated in a Gaussian Mixture Model (GMM) based speaker recognition system in [14]. LP-residual and pitch F_0 are considered as additional features.

The experimental results have shown that adding the pitch information has significant improvement when the correlation between the pitch and the cepstral coefficients is used. The combination of both the pitch F_0 and LPC residual features is proven to be effective. This approach achieved 98.5% speaker identification rate using NTT database.

PLAR (perceptual log area ratio), a new feature set for speaker identification task is introduced [15]. PLAR is closely related to the log area ratio (LAR) feature. PLAR is derived from the PLP (perceptual linear prediction) rather than the linear predictive coding (LPC). The PLAR feature derived from PLP is more robust to noise than LAR feature. PLAR, LAR and MFCC features were tested in a Gaussian mixture model (GMM) based speaker identification system. The F-ratio feature analysis has shown that the lower order PLAR and LAR coefficients are greater in classification performance to their MFCC counterparts. They had reported 98.81 % of speaker identification rate using TIMIT database.

Features derived from the Fourier transform phase have been explored [16]. The Modified Group Delay Feature (MODGDF) which is parameterized form of the modified group delay function is used as a front-end feature. They had developed a Gaussian mixture model (GMM) based speaker identification system with MODGDF as the front-end feature. This particular system is also tested on both clean (TIMIT) and noisy telephone (NTIMIT) speech.

A small set of low-level acoustic parameters that capture information about the speaker's source, vocal tract size and vocal tract shape is focused in [17]. It is evident that the set of eight acoustic parameters has comparable performance to the standard sets of 26 or 39 MFCCs for the speaker identification task. Gaussian mixture models were employed for building speaker models.

In the prototype proposed [18] for speaker identification system using auto-associate neural network (AANN) and formant features. The experiments demonstrate that formants extracted from difference spectrum perform significantly better than formant extracted from normal spectrum for the task of speaker identification. It has further demonstrated that formants from difference spectrum provide comparable speaker identification performance with that of features such as weighted linear predictive cepstral coefficients (LPCC) and Mel-Frequency Cepstral coefficients (MFCC). They had reported 100 % identification results on a data set of 50 speakers. The method proposed in [19], integrated the MFCC and phase information for speaker recognition. The speaker identification experiments were performed using NTT database and obtained the error reduction rate of 44.2 % than traditional MFCC.

The authors of [20], conducted different experimental studies on excitation component of speech for speaker recognition. Excitation information in speech is other form of LP-residual. AANN (Auto Associative Neural Network) models are used to capture speaker-specific information from the speech signal. It is claimed that AANN models can capture some speaker-specific excitation and needs significantly less amount of data both during training as well as testing, compared to the speaker recognition system using vocal tract information.

The author of [21] has reported, source features are used for automatic text-independent speaker recognition. He has illustrated that the source features are less prone to channel characteristics and noise. It is reported that the duration required to test the speaker model is 1 sec..

4. SPEAKER (SOURCE) CHARACTERISTICS IN THE LP RESIDUAL

Speech signal is produced as a resultant of interaction between the glottal excitation and vocal tract system. A simple block diagram representation of the speech production mechanism is shown in the Figure 2. Vibrations of the vocal folds, powered by air coming from the lungs during exhalation, are the sound source for speech. Hence, as can be from Figure 2, the glottal excitation forms the source, and the vocal tract forms the system. One of the prominent speech analysis techniques is the method of linear prediction analysis. The philosophy of linear prediction is closely associated with the basic speech production model. The LPC analysis approach performs spectral analysis on short segments of speech with an all-pole modeling constraint [20]. Because of speech can be modeled as the output of linear, time-varying system excited by a source, LPC analysis captures the vocal tract system information in terms of coefficients of the filter representing the vocal tract mechanism. Hence, analysis of speech signal by LP results in two components, namely the synthesis filter on one hand and the residual on the other hand. In brief, the LP residual signal is generated as a by product of the LPC analysis, and the computation of the residual signal is given below.

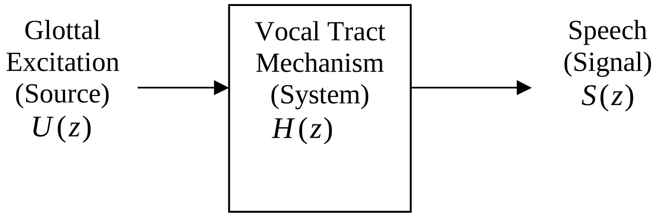

Figure 2. Source and System representation of speech production mechanism

If the input signal is represented by u_n and the output signal by s_n, then the transfer of the system can be expressed as, $H(z)=\dfrac{S(z)}{U(z)}$ (2)

Where $s(z)$ and $u(z)$ are z-transforms of s_n and u_n respectively.
Input signal can be computed with ensuing system and the output signal. The above equation can be expressed as $S(z) = H(z).U(z)$

$$U(z) = \frac{S(z)}{H(z)} \tag{3}$$

$$U(z) = \frac{1}{H(z)}.S(z) \tag{4}$$

$$U(z) = A(z).S(z) \tag{5}$$

Here $A(z) = {}^{1}\!/_{H(z)}$ is the inverse Z-transform which gives filter representation of the vocal tract system.

Linear prediction models the output s_n as the linear function of past outputs and present and past inputs. Since prediction is done by a linear function, the name linear prediction. Assuming an all-pole for the vocal tract, the signal s_n can be expressed as linear combination of past values and some input u_n as shown below.

$$Sn = -\sum_{k=1}^{p} a_k S_{n-k} + GU_n \tag{6}$$

Where G is a gain factor.

Now assuming that the input u_n is unknown, the signal s_n can be predicted only approximately from a linear weighted sum of past samples. Let this approximation of s_n be, where

$$\tilde{S}_n = -\sum_{k=1}^{p} a_k s_{n-k} \tag{7}$$

The difference of the error between the actual value S_n and predicted value is given as $e_n = S_n - \tilde{S}_n$ [22]. This error is nothing but LP residual of signal is shown in Figure 3.

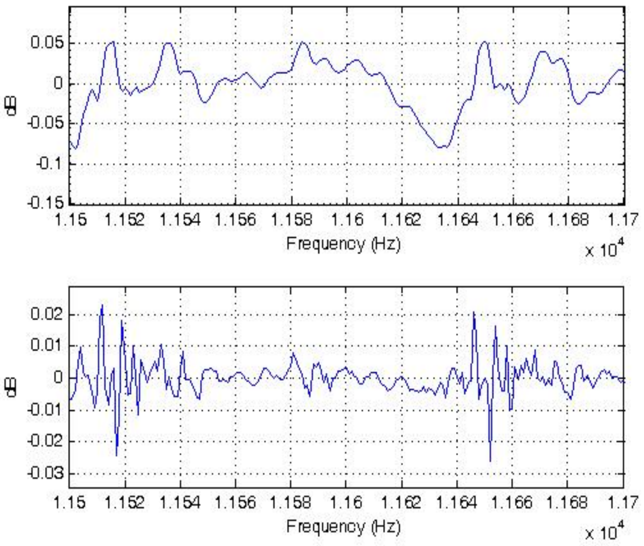

Figure 3. Actual signal and its LP residual

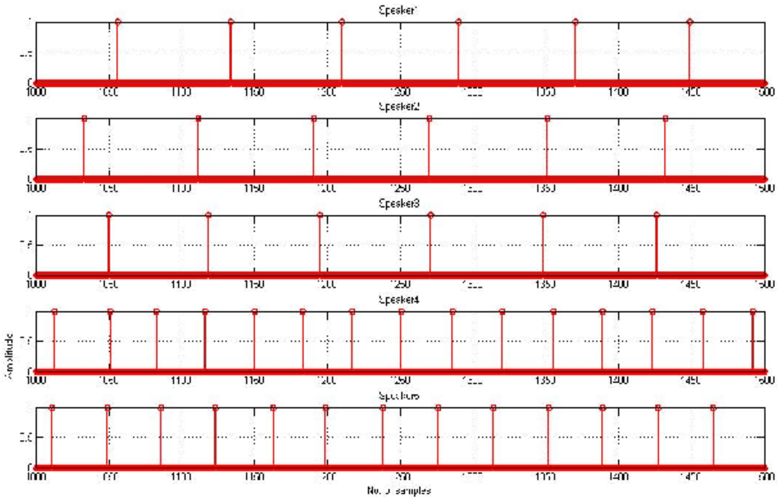

Figure 4. Instants of significant excitations for five male speakers for the sound unit /aa/

The significance of the source feature is illustrated in the Figure 4. The speech utterances sampled at 8 kHz were collected from five male speakers over a microphone. All the speakers uttered the sound unit /aa/. The significant instants of the glottal excitation are computed for the five speakers. The instants corresponding to the steady section of the utterances were displayed in the Figure 4. It is clearly seen from the Figure 4 that the periodicity of the instants of glottal excitation for each of the five speakers is different from that of the other’s. As shown in the Figure 4, it is clearly evident that the source features for different speaker is different.

5. Comparison of Speaker Specific Features with other Existing Feature Extraction Techniques

- Speaker specific information such as prosody [24], intonation, and duration are effectively captured in source features.
- Source features are less prone to channel characteristics and noise than other existing feature extraction techniques.
- For modeling, source information required is less than other existing techniques.

6. Conclusion

In this paper, we have discussed the different existing feature extraction techniques. We have illustrated source features for speaker indexing system, and how source features are unique than other existing systems. Finally, we propose source feature for robust speaker indexing system, which are less prone to channel characteristics and noise.

References

[1] J.P. Campbell (1997) JR., “Speaker recognition: a tutorial”, Proceedings of IEEE, vol. 85(9):1437-1462.

[2] Bonastre JF, Delacourt P, Fredouille C, Merlin T, and Wellekens C (2000) “A speaker tracking system based on speaker turn detection for NIST evaluation”, Proceedings of IEEE International Conference on acoustics, speech, and signal processing, pp. 1177–1180.

[3] Roy D, and Malamud C (1997) “Speaker identification based text to audio alignment for an audio retrieval system”, Proceedings of IEEE International Conference on acoustics, speech, and signal processing, pp. 1099–1102.

[4] Kimber DG, Wilcox LD, Chen FR, and Moran TP (1995) “Speaker segmentation for browsing recorded audio”, In: ACM CHI’95 Mosaic of Creativity, pp. 212–213.

[5] Mori, K. and Nakagawa, S. (2001) “Speaker change detection and speaker clustering using VQ distortion for broadcast news speech recognition”, Proc. IEEE International Conference on Acoustics, Speech and Signal Processing, Vol. 1, Salt Lake City, USA, pp. 413-416.

[6] J. J. Wolf, “Efficient acoustic parameters for speaker recognition”, J. Acoust. Soc. Amer., vol. 52, no. 6, pp. 2044-2056, 1972.

[7] M. Sambur, “Selection of acoustic features for speaker identification”, IEEE Transactions On ASSP, vol. 23, no. 2, pp. 176-182, 1975.

[8] S. Pruzansky and M. V. Mathews, “Talker-recognition procedure based on analysis of variance”, J. Acoust. Soc. Amer., vol. 36, no. 11, pp. 2041-2047, 1964.

[9] W. S. Mohn, “Two statistical feature evaluation techniques applied to speaker recognition”, IEEE Transactions on Computers, vol. 20, pp. 979-987, September 1971.

[10] N. Ney and R. Gierloff, “Speaker recognition using a feature weighting technique”, Proceedings of IEEE International Conference on acoustics, Speech, and Signal Processing, pp. 1645-1648, 1982.

[11] Y. Tohkura, “A weighted cepstral distance measure for speech recognition”, Proceedings of IEEE International Conference on acoustics, Speech and Signal Processing, vol. 11, pp. 761-764, 1986.

[12] Y. Tohkura, “A weighted cepstral distance measure for speech recognition”, IEEE Trans. Acoust., Speech, Signal Processing, vol. 35, no. 10, pp. 1414-1422, 1987.

[13] H.S. Jayanna, and S.R.M. Prasanna, “Analysis, feature extraction, modeling and testing techniques for speaker recognition”, IETE Technical Review, 2009, Vol. 26, issue 3, pp. 181-190, 2009.

[14] K.P. Markov and S. Nakagawa, “Integrating pitch and LPC-residual information with LPC-cepstrum for text-independent speaker recognition”, Jour. ASJ (E), Vol.20, no. 4, pp. 281–291, 1999.

[15] Chow D. and Abdulla W. H., “Robust Speaker identification Based on Perceptual Log Area Ration and Gaussian Mixture Models”, Proceedings of International Conference on Spoken Language Processing, 2004.

[16] Rajesh M. Hegde, Hema A. Murthy, and Gadde V. Ramana Rao, “ Application of the modified group delay function to speaker identification and discrimination”, in IEEE Trans. Acoust. Speech, Signal Processing, pp. 517-520, 2004.

[17] Carol Y. Espy-Wilson, Sandeep M., and Srikanth V., “ A New set of features for text- independent Speaker Identification”, Proceedings International Conference on Spoken Language Processing, pp. 1475-1478, 2006.

[18] Kishore Prahallad and et al., “Significance of formants from difference spectrum for speaker identification”, Proceedings International Conference on Spoken Language Processing, pp. 905-908, 2006.

[19] Seiichi Nakagawa, Kouhei Asakawa, and Longbiao Wang, “Speaker Recognition by Combining MFCC and phase information”, Proceedings International Conference on Spoken Language Processing, pp. 2005-2008, 2007.

[20] S. R. M. Prasanna, Cheedella S. Gupta, and B. Yegnanarayana, “Extraction of speaker-specific excitation information from linear prediction residual of speech”, Speech comm., vol. 48, pp. 1243-1261, June 2006.

[21] R.Rajeswara Rao, “Automatic text-independent speaker recognition using source feature”, Ph. D thesis, JNTUH.

[22] Molau S., Pitz M., Schluter R., and Ney H., “Computing mel-frequency cepstral coefficients on the power spectrum”, Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), vol. 1, pp. 73-76, May 2001.

[23] Padmanabhan M, Bahl LR, Nahamoo D, Picheny MA (1998) “Speaker clustering and transformation for speaker adaptation in speech recognition systems”, IEEE Transaction on Speech Audio Process 10:19-41.

[24] E. Shriberg, L. Ferrer , S. Kajarekar, A. Venkataraman, “Modeling prosodic feature sequences for speaker recognition “, Science Direct, Speech Communication, 2005.

Authors

Mr.V.Subba Ramaiah received his B.Tech. degree in Computer Science and engineering from SITAMS, JNT University, Chittoor, India, in 2002 and the M.Tech. degree in Computer Science from SIT, JNT University, Hyderabad, India, in 2007. He has been working as Senior Assistant Professor in the department of Computer Science and Engineering, Mahatma Gandhi Institute of Technology, Hyderabad. His research interests are speech and pattern recognition.

Dr. R.Rajeswara Rao was born in India in 1976. He received his B.Tech. degree in Computer Science and engineering from Siddhartha Engineering College, Vijayawada, India, in 1999 and the M.Tech. degree in Computer Science and Engineering from College of Engineering, JNT University, Hyderabad, India, in 2003. He has completed his Ph.D degree in computer science and engineering from JNT University, Hyderabad, India, in 2010. He is currently Professor and Head in the Department of Computer Science and engineering, Mahatma Gandhi Institute of Technology, Hyderabad. His research interests are speech processing and pattern recognition.