International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 3 Issue 12, December-2014

Prosody Based Speaker Verification System: Effect of Voice Disguise

Sophia Susan John, Leena Mary, Anil P Antony
Dept. of Electronics and Communication Engineering
Rajiv Gandhi Institute of Technology
Kerala, India

***Abstract*—Speaker recognition systems based on spectral features perform well in acoustically matched and noise-free conditions. However, they fail to model information about the speaker at many other levels that might contribute to speaker recognition. In this work, we have studied the effects of voice disguise on prosody based speaker verification system. For this a prosody based speaker verification system is implemented using support vector machine (SVM) modeling. For the extraction of prosodic features speech is segmented into phrase-like regions by detecting long pauses regions and further segmentation is done at the valleys of short time energy contour. Then prosodic features such as intonation, duration and intensity features are extracted and used for modeling. Though this prosody based speaker model works well, the system fails in some cases due to choice universal hard threshold value for verification. Therefore a set of impostor models were created for male and female speakers. During testing phase gender was identified using average pitch values to select the appropriate set of background models, which helped to fix the threshold adaptively and this system gave improved performance. In case of voice disguises, it was observed that the number of feature vectors obtained were minimal as per the existing segmentation algorithm. Therefore segmentation algorithm was modified to give more meaningful syllable-like segmentation resulting better performance of the prosody-based speaker verification.**

Keywords—automatic speaker verification; prosodic features; prosody; voice disguise; syllabification

I. INTRODUCTION

The goal of automatic speaker recognition systems is to extract, characterize and recognize the information in the speech signal conveying speaker identity. Each person has unique anatomy, physiology and learned habits that familiar persons use in everyday life to recognize the person. State-of-the-art speaker recognition systems use a number of these features in parallel, attempting to cover these different aspects and employing them in a complementary way to achieve more accurate recognition [1].

Robustness of automatic speaker recognition is critical for real-world applications. With recent widespread use of speech-enabled services for consumers and growing importance of speaker recognition in security and defense, it is important to have a robust system.

Work described in this paper is financially supported by Kerala State Council for Science, technology and Environment (KSCSTE), Government of Kerala, India.

Although current state-of the-art speaker verification systems achieve very high performance on clean data, there are few studies of noisy conditions. So noise robust speaker verification systems are of great importance. In daily acoustic environments, additive noise, room reverberation and channel/handset variations conspire to pose considerable challenges to such systems. A lot of research has been devoted to dealing with individual challenges. For example, speakers can be modeled in multiple noisy environments to reduce the mismatch between training and test conditions [2].

Current speaker recognition systems are dominated by the vocal tract characteristics represented using spectral/cepstral features such as MFCC (mel-frequency cepstral coefficients) and LPCC (linear prediction cepstral coefficients) derived through short-time spectral analysis. Such systems perform well in acoustically matched and noise-free conditions [1].

However, they fail to model information about the speaker at many other levels that might contribute to speaker recognition. Spectral features are affected by environment variability, channel and noise. This is due to, the characteristics of speech signal are influenced by the environment in which it is collected and channel through which it is transmitted. These factors can significantly change the features derived through short-time spectral analysis [3]. Therefore it is important to have features which are less affected by channel and environment characteristics for a robust recognition system. The prosodic features derived from pitch contour, amplitude contour, duration etc. are less affected by channel characteristics, environment and noise [1].

Voice disguise is considered as a deliberate action of a speaker who wants to falsify or conceal his identity [4]. A disguise is applied when there is a deliberate will to transform one's voice to imitate someone or just to change the sound. It may not be possible for an impersonator to imitate all characteristics linked to prosody such as pitch register, voice quality, dialect or the speech style of the target speaker's voice, but some specific parameters are enough to disturb the recognition.

In the field of automatic speaker recognition application, one of the most threats is voice disguise. Mimicking is one form of voice disguise where a person modifies his voice to sound like another person [4]. The speaker could transform his voice by electronic scrambling or more simply by exploiting the intra-speaker variability: modification of his own pitch, modification of the position of the articulators like lips or tongue which affect the formant frequencies.

Major application of automatic speaker recognition is in remote access control such as online banking, e-commerce and other consumer related applications. To use speech as a biometric in such applications, performance of speaker recognition should be studied in the context of mimicked speech [5]. The automatic speaker recognition, which relies on spectral features, suffers from the problem of lack of robustness and interpretability for forensic applications [6]. Robustness of prosodic features in cases of acoustic mismatch may provide an edge for them in forensic applications. So for speaker recognition purpose optimum features should robust against disguise.

There are some characteristics that lends naturalness to speech. The variation of pitch provides some recognizable melodic properties to speech. This controlled modulation of pitch is referred as intonation. The sound units are shortened or lengthened in accordance to some underlying pattern giving rhythmic properties to speech. Some syllables or words may be made more prominent than others, resulting in linguistic stress. Prosodic characteristics such as intonation, rhythm and stress in speech convey some important information regarding the identity of the speaker [1]. These characteristics are called prosodic cues. Each prosodic cue is expressed using three acoustic parameters: pitch, duration and energy.

The remaining part of this paper is structured as follows. Section II presents the prosody based speaker verification system. Section III explains the effect of voice disguise on prosody based speaker verification system with modified segmentation method. In Section IV, the performance evaluations on the system with original speech and mimicked speech are described.

I. PROSODY BASED SPEAKER VERIFICATION SYSTEM

In our previous work, we implemented a simple prosody-based speaker verification system described in [7]. The system operates in two phases: training and verification phases. In the training phase new speaker (with known identity) is enrolled into the system's database. In the verification phase, an unknown speaker gives a speech input along with a claim and the system makes a decision to accept/reject speaker's claim. The system fails to accept the genuine speaker and does not reject the mimicked speech in some of the cases. Therefore it is necessary to modify the system.

In this work, an automatic speaker verification system is implemented, as shown in Fig. 1, to study the effects voice disguise on the system. The implemented verification system is text independent. For this work two categories of speech data is used. One is speech of the selected celebrities and the other is speech of professional mimicry artists while imitating selected celebrities. The first category of speech (of popular film stars and politicians) was collected from Internet. The second category was recorded in the laboratory environment.

Both training and verification phases include feature extraction. The feature extractor converts the digital speech signal into a sequence of numerical descriptors, called feature vectors. The prosodic features are extracted and represented using the segmentation method [1]. This feature vector consists of the following prosodic features:

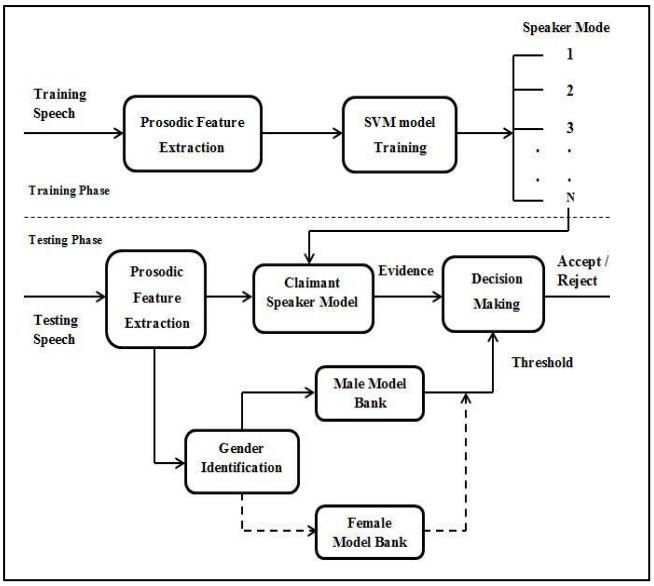

Fig. 1. Training and testing phase of speaker verification system using modified threshold.

- Four weights of Legendre polynomial.
- Jitter (the cycle-to-cycle variation of fundamental frequency).
- Delta energy (the difference between maximum energy and minimum energy in the frame.
- Shimmer (which is expressed as the variability of the peak-to-peak amplitude.)
- Total duration and voiced duration.

A. Segmentation

In the ASR (Automatic Speech Recognizer) free approach, segment boundaries are estimated using cues derived from the speech signal [1] [4]. For representing syllable-based rhythm, intonation, and stress, the speech signal should be segmented.

Speech may be segmented into phrase-like regions using fairly long pauses separating them. For this, speech/non-speech classification is performed and first level of segmentation is performed based on this. Further segmentation is done using the valleys in the short term energy. Prosodic features representing intonation, duration and energy are then extracted from these segments [4]. One of the basic short-time features useful for speech/non-speech classification is the short-time energy. The short-time energy, E_n is defined as,

$$E_n = \sum_{m=1}^{L} (x[m]w[n-m])^2 \quad (1)$$

where L is the number of samples of the speech signal, $w[n\text{-}m]$ represents a time shifted window sequence, whose purpose is to select a segment of the sequence $x[m]$ in the neighborhood of sample $m = n$.

The other useful feature for speech/non-speech classification is the most dominant frequency component of the speech frame spectrum. The most dominant frequency (MDF) is the maximum value of the spectrum magnitude [4]. The last feature extracted is the voicing information from pitch. For each frame, speech/non-speech decision is taken based on whether the feature has value greater than a threshold and whether the frame is voiced/unvoiced.

B. *SVM based speaker modeling*

The speaker and the reference population are modeled using SVM [8]. The first step of SVM training is transform data to the format of an SVM package. Then scaling of data is done. The main advantage of scaling is to avoid attributes in greater numeric ranges dominating those in smaller numeric ranges. Another advantage is to avoid numerical difficulties during the calculation.

SVM constructs non-linear decision regions (hyper plane) in discriminative manner. In this work, SVM is implemented using the LIB-SVM [9]. The extracted features are arranged in the format, which is supported by the LIB-SVM classifier. The kernel used in this work is Radial Basis Function (RBF). In order to optimize the classification performance the best values for error penalty parameter (c) and kernel parameter (γ) are found using grid search algorithm. The best values of the parameters c and γ are used to training the whole training set.

During training phase SVM models are created for each celebrity using prosodic features derived from their original (training) speech. These models are tested using prosodic features extracted from the given test speech (includes original and mimicked speech). In this work, a set of impostor models were created for male and female speakers.

Decision was made using the match score or evidence value obtained with respect to a threshold. The binary decision is either acceptance or rejection of the speaker. During testing the evidence value is obtained from the claimant model using SVM prediction. The threshold is selected as the mean score of background (gender-specific) models, in order to make it speaker-dependent.

II. Effect of Voice Disguise on Prosody-based Speaker Verification System

The principle of voice disguise is based on the impersonation of some specific characteristics of a target voice linked to prosody, pitch register, voice quality, dialect or the speech style. In theory, a subject could modify prosodic parameters such as intonation, loudness and rhythm by changing pitch, intensity and duration to mimic the speaking style of another person. At the production level, these changes are brought out by varying the vocal fold tension, sub-glottal pressure and airflow, to change pitch, intensity and duration respectively [5].

Table I. shows the performance evaluation of prosody based speaker verification system with original speech and mimicked speech, when threshold was set as mean value obtained from background models for decision making. The result shows, the overall performance of the system was poor.

A. *Modification of Threshold for Decision Making*

Though the prosody-based speaker model works well, the system fails in some cases. Therefore a set of separate impostor models were created for male and female speakers. During testing phase gender was identified using average pitch values to select the appropriate set of background (male/female) models, which helped to fix the threshold adaptively. Fig. 1 shows prosody-based speaker verification system with modified threshold. In order to improve the accuracy the system, we modified the threshold as *2*standard deviation + mean* for decision making, which was obtained from background models.

B. *Modified Segmentation Method for the Extraction of Prosodic Features*

In the previous segmentation method, as explained in section II, segmentation is done at the valleys in the short term energy. The prosodic features are extracted from these segments. Sometimes these segments can be large in duration leading less number of feature examples. This may lead to reduced performance of the system.

For first level of segmentation, here also speech/non-speech classification is used. Further segmentation is done by syllabification using valley detection and vowel onset point (VOP) [10] [11]. By using modified segmentation method, number of feature vectors obtained was nearly doubled, which may lead to better speaker modeling. Here the segments are syllables which are linguistically meaningful for representing prosodic features.

III. Performance Evaluation

In this section, evaluation results on the prosody-based speaker verification system are discussed. In all evaluation cases, 11-15 minutes speech data was used for training from each celebrity and 3-5 minutes speech data (original and mimicked speech from each selected celebrity) was used for testing.

Table I. shows the performance evaluation of prosody based speaker verification system when threshold was set as mean value obtained from background models for decision making. Table shows, the system overall performance was low for all models. Also it was observed that there is a variation in mean pitch value that could affect the accurate recognition of speaker verification system.

Table II. shows the overall performance of this prosody-based speaker verification system with modified threshold. It shows the efficiency of the system could be improved to some extent, but the system fails in some case due to the effects of voice disguise. Here, original speech of selected celebrities was collected from Internet but speech of professional mimicry artists while imitating selected celebrities were recorded in the laboratory environment. So sources of data used are different and it was observed that this could affect the system performance.

Table III. shows the performance evaluation (in percentage) of the system with modified segmentation method.

TABLE I. PROSODY-BASED SPEAKER VERIFICATION SYSTEM: PERFORMANCE EVALUATION USING ORIGINAL AND MIMICKED SPEECH

Speaker Model	Performance in Percentage		Overall Performance in Percentage
	Original Speech	*Mimicked Speech*	
Celebrity 1	60	70	65
Celebrity 2	70	20	45
Celebrity 3	63	50	56.5

TABLE II. PROSODY-BASED SPEAKER VERIFICATION SYSTEM: PERFORMANCE EVALUATION WITH MODIFIED THRESHOLD USING ORIGINAL AND MIMICKED SPEECH

Speaker Model	Performance in Percentage		Overall Performance in Percentage
	Original Speech	*Mimicked Speech*	
Celebrity 1	80	70	75
Celebrity 2	90	30	60
Celebrity 3	87.5	60	73.75

TABLE III. PROSODY-BASED SPEAKER VERIFICATION SYSTEM WITH MODIFIED SEGMENTATION METHOD: PERFORMANCE EVALUATION USING ORIGINAL AND MIMICKED SPEECH

Speaker Model	Performance in Percentage		Overall Performance in Percentage
	Original Speech	Mimicked Speech	
Celebrity 1	90	83	86.5
Celebrity 2	100	70	85
Celebrity 3	100	75	87.5

The system performance could be improved with modified segmentation method. In the case of prosody, speech data with more duration is required for the extraction of more feature examples. Here also it was observed that the variation in mean pitch value, which affects the gender identification and thus the selection of model bank.

IV. CONCLUSION

In this paper, effects of voice disguise on the prosody-based speaker verification system are described. The prosodic features are extracted and represented using segmentation. During testing phase gender was identified using average pitch values to select the appropriate set of background models, which helped to fix the threshold adaptively and this system gave improved performance. In case of voice disguises, it was observed that the number of feature vectors obtained were minimal as per the existing segmentation algorithm. Therefore segmentation algorithm was modified to give more meaningful syllable-like segmentation resulting better performance of the prosody-based speaker verification. In the future work, a set of prosodic features should be identified which are less affected by channel and noise. It is necessary to suggest a pitch extraction method which works well for speech signal with channel mismatch and noise.

ACKNOWLEDGMENT

The authors would like to thank Mr. Jubin James Thennattil of ADSPR Lab, Rajiv Gandhi Institute of Technology, Kottayam for his support.

REFERENCES

[1] L. Mary and B. Yegnanarayana, "Extraction and representation of prosodic features for language and speaker recognition," Science Direct, speech communication, vol. 50, no. 10, pp. 782-796, 2008.

[2] B. Yegnanarayana and S. Kishore, " Robust speaker identification in noisy and reverberant conditions," IEEE/ACM Transactions on audio, speech and language processing, vol. 22, no. 4, pp. 836-845, 2014.

[3] L. R. Rabiner, Biing Hwang Juang, and B. Yegnanarayana, Fundamentals of Speech Recognition,in Magnetism, 1st ed., Pearson 2009.

[4] L. Mary, A. Babu, and A. Joseph, "Analysis and detection of mimicked speech based on prosodic features," in International Journal of Speech Technology, vol. 15, no. 3, pp. 107-117, 2012.

[5] M. Blomberg, D. Elenius and E. Zetterholm, "Speaker verification scores and acoustic analysis of a professional impersonator," in Proceedings FONETIK 2004, the XVIIth Swedish phonetics conference, pp. 84-87, 2004.

[6] E. Shriberg and Stolcke, "The case for automatic higher level features in forensic speaker recognition," in Proceedings of inter speech, pp. 1509-1512, 2008.

[7] Anil P. Antony, Sophia Susan John and L. Mary, "Automatic Speaker Verification using Prosodic Features" in National Conference on Technological Trends (NCTT 2014), pp. 1454-1457, Aug. 2014.

[8] W. Campbell, J. Campbell, D. Reynolds, E. Singer, and P. Torres-Carrasquillonil, "Support vector machines for speaker and language recognition" in Comput. Speech Lang., vol. 20, pp. 210-229, 2006.

[9] J. Chih-Chung. Chang and L. Chih-Jen. Lin, "LIBSVM : a library for support vector machines," in CM Transactions on Intelligent Systems and Technology, Software available at http://www.csie.ntu.edu.tw/cjlin/libsvm.

[10] S. R. Mahadeva Prasanna and B. Yegnanarayana, "Detection of vowel onset point events using excitation information," in Proceedings of interspeech, pp. 1133-1136, 2005.

[11] M. R. Gayathri, Anil P. Antony, L. Mary, "Syllabification of Speech Signals," in National Conference on Technological Trends (NCTT 2014), pp. 1448-1453, Aug. 2014.