SPEAKER VERIFICATION USING ACOUSTIC AND PROSODIC FEATURES

Utpal Bhattacharjee[1] and Kshirod Sarmah[2]

Department of Computer Science and Engineering, Rajiv Gandhi University, Rono Hills, Doimukh, Arunachal Pradesh, India, PIN – 791112

[1]utpalbhattacharjee@rediffmail.com

[2]kshirodsarmah@gmail.com

ABSTRACT

In this paper we report the experiment carried out on recently collected speaker recognition database namely Arunachali Language Speech Database (ALS-DB)to make a comparative study on the performance of acoustic and prosodic features for speaker verification task.The speech database consists of speech data recorded from 200 speakers with Arunachali languages of North-East India as mother tongue. The collected database is evaluated using Gaussian mixture model-Universal Background Model (GMM-UBM) based speaker verification system. The acoustic feature considered in the present study is Mel-Frequency Cepstral Coefficients (MFCC) along with its derivatives.The performance of the system has been evaluated for both acoustic feature and prosodic feature individually as well as in combination.It has been observed that acoustic feature, when considered individually, provide better performance compared to prosodic features. However, if prosodic features are combined with acoustic feature, performance of the system outperforms both the systems where the features are considered individually. There is a nearly 5% improvement in recognition accuracy with respect to the system where acoustic features are considered individually and nearly 20% improvement with respect to the system where only prosodic features are considered.

KEYWORDS

Speaker Verification, Multilingual,GMM-UBM,MFCC, Prosodic Feature

1. INTRODUCTION

Automatic Speaker Recognition (ASR)refers to recognizing persons from their voice. The sound of each speaker is not identical because their vocal tract shapes, larynx sizes and other parts of their voice production organs are different. ASR System can be divided into either Automatic Speaker Verification(ASV) or Automatic Speaker Identification (ASI) systems [1,2, 3].Speaker verification aims to verify whether an input speech corresponds to the claimed identity. Speaker identification aims to identify an input speech by selecting one model from a set of enrolled speaker models: in some cases, speaker verification will follow speaker identification in order to validate the identification result [4].Speaker Verification is the task of determining whether a person is who he or she claims to be (a yes/ no decision).Since it is generally assumed that imposter (falsely claimed speaker) are not known to the system, it is also referred to as an Open-Set task [5].

The speaker verification system aims to verify whether an input speech corresponds to the claimed identity or not. A security system based on this ability has great potential in several application domains. Speaker verification systems are typically distinguished into two categories – text-dependent and text-independent [6]. In text-dependent system, a predetermined group of words or sentences is used to enroll the speaker to the system and those words or sentences are used to verify the speaker. Text-dependent system use an explicit verification protocol, usually combined with pass phrases or Personal Identification Number (PIN) as an

DOI : 10.5121/acij.2013.4105

additional level of security. In text-independent system, no constraints are placed on what can be said by the speaker. It is an implicit verification process where the verification is done while the user is performing some other tasks like talking with the customer care executive or registering acomplain.

The state-of-art speaker verification system use either adaptive Gaussian mixture model (GMM) [7] with universal background model (UBM) or support vector machine (SVM) over GMM super-vector [8].

Mel-frequency Cepstral coefficients (MFCCs) are most commonly used feature vector for speaker verification system. Supra-segmental features like – prosody, speaking style are also combined with the cepstral feature to improve the performance[9].

Prosody plays a key role in the perception of human speech. The information contained in prosodic features is partly different from the information contained in cepstral features. Therefore, more and more researchers from the speech recognition area are showing interests in prosodic features. Generally, prosody means "the structure that organizessound". Pitch (tone), Energy (loudness) and normalized duration (rhythm) are the main components of prosody for a speaker. Prosody can vary from speaker to speaker and relies on long-term information of speech.

Very often, prosodic features are extracted with larger frame size than acoustical features since prosodic features exist over a long speech segment such as syllables. The pitch and energy-contours change slowly compared to the spectrum, which implies that the variation can be captured over a long speech segment [10].

The rest of the paper is organized as follows: Section–2 describes the details of the speaker recognition database. Section–3 details the speaker verification system. The experimental setup, data used in the experiments and result obtained are described in Section 4. The paper is concluded in Section–5.

2. SPEAKER RECOGNITION DATABASE

In this section we describe the recently collected Arunachali Language Speech Database (ALS-DB).Arunachal Pradesh of North East India is one of the linguistically richest and most diverse regions in all of Asia, being home to at least thirty and possibly as many as fifty distinct languages in addition to innumerable dialects and subdialects thereof [11]. The vast majority of languages indigenous to modern-day Arunachal Pradesh belong to the Tibeto-Burman language family. The majority of these in turn belong to a single branch of Tibeto-Burman, namely Tani. Almost all Tani languages are indigenous to central Arunachal Pradesh while a handful of Tani languages are also spoken in Tibet. Tani languages are noticeably characterized by an overall relative uniformity, suggesting relatively recent origin and dispersal within their present-day area of concentration. Most Tani languages are mutually intelligible with at least one other Tani language, meaning that the area constitutes a dialect chain. In addition to these non-Indo-European languages, the Indo-European languages Assamese, Bengali, English, Nepali and especially Hindi are making strong inroads into Arunachal Pradesh primarily as a result of the primary education system in which classes are generally taught by immigrant teachers from Hindi-speaking parts of northern India. Because of the linguistic diversity of the region, English is the only official language recognized in the state.

To study the impact of language variability on speaker recognition task, ALS-DB is collected in multilingual environment. Each speaker is recorded for three different languages – English, Hindi and a local language, which belongs to any one of the four major Arunachali languages -

Adi, Nyishi, Galo and Apatani. Each recording is of 4-5 minutes duration. Speech data were recorded in parallel across four recording devices, which are listed in table -1.

Table 1: Device type and recording specifications

Device Sl. No	**Device Type**	**Sampling Rate**	**File Format**
Device 1	Table mounted microphone	16 kHz	wav
Device 2	Headset microphone	16 kHz	wav
Device 3	Laptop microphone	16 kHz	wav
Device 4	Portable Voice Recorder	44.1 kHz	mp3

The speakers are recorded for reading style of conversation. The speech data collection was done in laboratory environment with air conditioner, server and other equipments switched on. The speech data was contributed by 112 male and 88 female informants chosen from the age group 20-50 years. During recording, the subject was asked to read a story from the school book of duration 4-5 minutes in each language for twice and the second reading was considered for recording. Each informant participates in four recording sessions and there is a gap of at least one week between two sessions.

3. SPEAKER VERIFICATION SYSTEM

In this works, aSpeaker Verification system has been developed using Gaussian Mixture Model with Universal Background model (GMM-UBM) based modeling approach. A 39-dimensional feature vector was used, made up of 13 mel-frequency cepstral coefficient (MFCC) and their first order and second order derivatives. The first order derivatives were approximated over three samples and similarly for second order derivatives.The coefficients were extracted from a speech sampled at 16 KHz with 16 bits/sample resolution. A pre-emphasis filter$H(z) = 1 - 0.97z^{-1}$ has been applied before framing. The pre-emphasized speech signal is segmented into frame of 20 microseconds with frame frequency 100 Hz. Each frame is multiplied by a Hamming window. From the windowed frame, FFT has been computed and the magnitude spectrum is filtered with a bank of 22 triangular filters spaced on Mel-scale. The log-compressed filter outputs are converted to cepstral coefficients by DCT. The 0^{th}cepstral coefficient is not used in the cepstral feature vector since it corresponds to the energy of the whole frame [12], and only next 12 MFCC coefficients have been used. To capture the time varying nature of the speech signal, the MFCC coefficients were combined with its first order and second derivatives, we get a 39-dimensional feature vector.

Cepstral Mean Subtraction (CMS) has been applied on all features to reduce the effect of channel mismatch. In this approach we apply Cepstral Variance Normalization (CVN) which forces the feature vectors to follow a unit variance distribution in feature level solution to get more robustness results.

The prosodic features typically include pitch, intensity and normalized duration of the syllable. However, as a limitation of combination, the features have to be frame based and therefore only pitch and intensity are selected to represent the prosodic information in the present study. Pitch and intensity are static features as they are calculated frame by frame. They only represent the exact value of the current frame. In order to incorporate more temporal information their 1st order and 2^{nd} order derivatives has also been included.

The Gaussian mixture model with 1024 Gaussian components has been used for both the UBM and speaker model. The UBM was created by training the speaker model with 50 male and 50

female speaker's data with 512 Gaussian components each male and female model with Expectation Maximization (EM) algorithm. Finally UBM model is created by pulling the both male and female models and finding the average of all these models [13]. The speaker models were created by adapting only the mean parameters of the UBM using maximum a posteriori (MAP) approach with the speaker specific data [8].

The detection error trade-off (DET) curve has been plotted using log likelihood ratio between the claimed model and the UBM and the equal error rate (EER) obtained from the DET curve has been used as a measure for the performance of the speaker verification system. Another measurement Minimum DCF values has also been evaluated.

4. EXPERIMENTS AND RESULTS

All the experiments reported in this paper are carried out using the database ASL-DB described in section 2. An energy based silence detector is used to identify and discard the silence frames prior to feature extraction. Only data from the headset microphone has been considered in the present study. All the four available sessions were considered for the experiments. Each speaker model was trained using one complete session. The test sequences were extracted from the next three sessions. The training set consists of speech data of length 120 seconds per speaker. The test set consists of speech data of length 15 seconds, 30 seconds and 45 seconds. The test set contains more than 3500 test segments of varying length and each test segment will be evaluated against 11 hypothesized speakers of the same sex as segment speaker [14].

In this experiment single language (English, Hindi, and a Local language) has been considered for training the system and each language has been considered separately for testing the system. Training sample of length 120 seconds from a single session has been considered for training the system and the other three sessions have been considered for testing the system.

Figure–1(a),(b) and (c) shows the DET curves for Speaker Verification system obtained for the three languages in the speech database. The result of the experiments has been summarized in table–1.

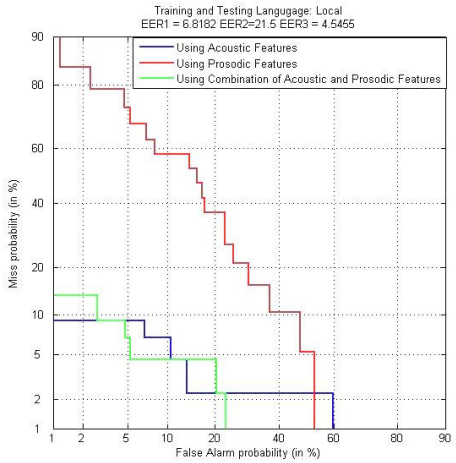

Figure 1(a). DET curves for the speaker verification system using the Local language as training and testing language

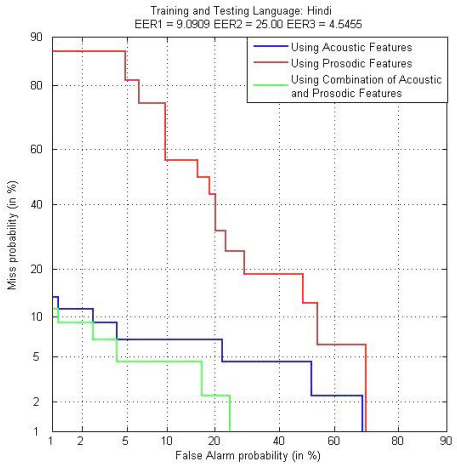

Figure 1(b). DET curves for the speaker verification system using the Hindi language as training and testing language

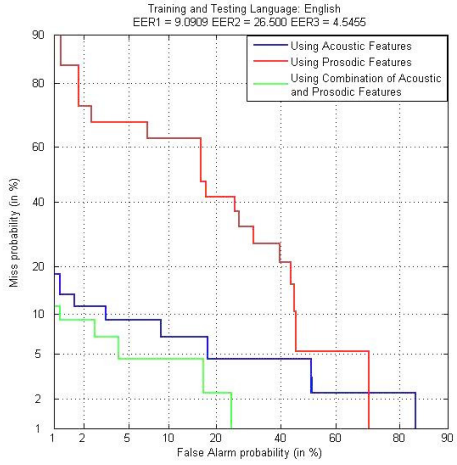

Figure 1(c). DET curves for the speaker verification system using the English language as training and testing language

Table 2.EER andMin DCF values for speaker verification system for training with one language and testing with each language

Training & Testing Language	Feature Vectors	ERR%	Recognition Rate%	Minimum DCF Value
Local	MFCC + ΔMFCC+ ΔΔMFCC	6.81	93.19	0.0991
	PROSODIC	21.50	79.50	0.4045
	MFCC + ΔMFCC+ ΔΔMFCC+ PROSODIC	4.55	95.45	0.0823
Hindi	MFCC + ΔMFCC+ ΔΔMFCC	6.81	93.19	0.0968
	PROSODIC	25.00	75.00	0.4438
	MFCC + ΔMFCC+ ΔΔMFCC+ PROSODIC	4.55	95.45	0.0927
English	MFCC + ΔMFCC+ ΔΔMFCC	9.09	90.91	0.1195
	PROSODIC	26.50	73.50	0.4632
	MFCC + ΔMFCC+ ΔΔMFCC+ PROSODIC	4.55	95.45	0.0823

5. CONCLUSIONS

The experiments reported in the above section established the fact that MFCC features along with its time derivatives may be considered as an efficient parameterization of the speech signal for speaker verification task. It has been observed that when MFCC with its 1st and 2nd order derivatives has been considered as feature vector, it gives a recognition accuracy of around 92.43%. Performance of another feature vector namely prosody has also been evaluated in the present study. It has been observed that when the prosodic features are considered as feature vector for the speaker verification system, it gives a recognition accuracy of 76%, which is nearly 16% below the recognition accuracy MFCC features vector. However, it has been observed that when both MFCC and prosodic features are combined, a recognition accuracy of 95.45% has been achieved. Thus, it may be conclude that MFCC features when combined with prosodic features, the performance of the system improves marginally, in the present study by 3%, without increasing the complexity of the system. Another interesting observation made in the present study is that when English languages has been considered for training and testing the system, a recognition accuracy of 90.19% has been achieved with MFCC features whereas under same experimental condition the recognition accuracy for Local and Hindi languages is 93.19%. The relatively poor performance in case of English can be explained in the context of linguistic scenario of Arunachal Pradesh. The people of Arunachal Pradesh of India, specially the educated section use Hindi in their day-to-day conversation even with their family members. Therefore, Hindi is as fluent to them as their native language and its articulation is relatively robust to context like their native language. However, English being the non-native language undergoes major deviation in the articulation with context.

ACKNOWLEDGEMENT

This work has been supported by the ongoing project grant No. 12(12)/2009-ESD sponsored by the Department of Information Technology, Government of India.

REFERENCES

[1] Furui, S., (1997) "Recent advances in speaker recognition", *Pattern Recognition Lett.*, 18, 859–872.

[2] Campbell, J.P., Jr., (1997) "Speaker recognition: a tutorial", Proceedings of the IEEE, 85(9): 1437–1462.

[3] Bimbot, F. et al, (2004), "A tutorial on text-independent speaker verification", *EURASIP Journal on Applied Signal Processing*, 4, 430–451.

[4] Park, A. and Hazen, T. J. (2002), "ASR dependent techniques for speaker identification", *Proceedings of the International Conference on Spoken Language Processing*, Denver, Colorado.

[5] Douglas A. Reynolds (2002), "An overview of Automatic Speaker Recognition Technology", MIT Lincoln Laboratory,244 wood St. Lexinton, MA 02140,USA,IEEE 2002.

[6] Rosenberg, J. Delong, C. Lee, B. Juang and F. Soong, (1992), "The use of cohort normalized scores for speakerrecognition," *InProc. ICSLP*, pp. 599–602.

[7] Reynolds, A. (1995), "Robust text-independent speakeridentification using Gaussian mixture speaker models," *Speech Communications*, vol. 17, pp. 91-108.

[8] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn (2000), "SpeakerVerification Using Adapted Gaussian Mixture Models," *Digital Signal Processing*, vol. 10(1–3), pp. 19-41.

[9] Haris B.C., Pradhan G., Misra A, Shukla S., Sinha R and Prasanna S.R.M. (2011), Multi-variability Speech Database for Robust Speaker Recognition, *In Proc. NCC*, pp. 1-5.

[10] Shriberg, E.E., (2007), "Higher Level Features in Speaker Recognition", *In C. Muller (Ed.) Speaker Classification I. Volume 4343 of Lecture Notes in Computer Science / Artificial Intelligence. Springer: Heidelberg / Berlin / New York*, pp. 241-259.

[11] Arunachal Pradesh, http://en.wikipedia.org/wiki/Arunachal_Pradesh.

[12] Xiaojia Z., Yang S.andDeLiang W., (2011), Robust speaker identification using a CASA front-end, *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*, pp.5468-5471, 2011.

[13] Kleynhans N.T. and Barnard E., (2005), Language dependence in multilingual speaker verification, *In Proc. of the 16th Annual Symposium of the Pattern Recognition Association of South Africa*, Langebaan, South Africa, pp. 117-122.

[14] NIST 2003 Evaluation plan, http:// www.itl.nist.gov/iad/mig/tests/sre/2003/2003-spkrec-evalplan-v2.2.