

Prosodic-spectral feature fusion for robust speaker recognition on telephone speech

Brodwyn L. Appanna, Marshalleno Skosan, Daniel J. Mashao

Department of Electrical Engineering
University of Cape Town, Rondebosch, 7700, South Africa
bappanna@crg.ee.uct.ac.za mskosan@crg.ee.uct.ac.za daniel@eng.uct.ac.za

Index Terms—speaker recognition, telephone speech, feature fusion, prosodic features

Abstract—As the digital divide between man and information decreases, personal information via public media, e.g. the telephone, will become more accessible. Consequently, there will be a greater demand on reliable, robust speaker recognition systems and thus a demand on exploration of different paths within the research. One of the relative recent paths has been in exploring different types of feature sets, viz. high-level features. Traditional and current speaker recognition systems primarily use low-level (physiological) features of speech that model the physical dimensions of the vocal tract. The popular MFCC is such a feature vector. There is a growing trend in the literature, however, that evidently supports the idea of improved systems by fusing low-level features with high-level (psychological, learned) features like conversational, lexical, phonemic and prosodic patterns found in speech. This work, which is part of the wider study of many high-level low-level feature fusion systems, investigated the performance of a speaker ID system evaluated on the NTIMIT database. The system employs the popular MFCC feature vector concatenated with a single simple, low-end high-level feature vector containing prosodic information. The vector, developed by Wildermoth and Paliwal, contains the maximum autocorrelation values of a segmented frame of speech and is accordingly named the MACV feature. Results presented in this paper showed an improvement from 82.74% to 85.32% for the fused system, a relative improvement of over 3% for the identification rate. The increase in performance on a popular, state-of-the-art feature vector, like the MFCC, creates anticipation for promising results to future work on similar systems used on more challenging databases with more complex high-level feature sets.

I. INTRODUCTION

“There are two main sources of speaker-specific characteristics of speech: physical and learned” [1]. The former is based on the alteration of an acoustic wave’s frequency content as it passes through the vocal tract. The resonances of the vocal tract (formants), determined by its physical dimensions, modifies the acoustic wave’s spectrum [2], [3]. This feature is what we perceive as the sound of a voice. The second source of speaker-specific characteristics are psychological or habitual rather than physiological. They include features like conversational, lexical, phonemic and prosodic patterns of speech [3].

Reynolds, et. al. [4] describes the characteristics of speech as existing in a hierarchy that increases with complexity from bottom to top. Physical features are found at the bottom

and psychological features above, with different psychological features existing at different levels of the hierarchy. Speaker recognition systems make use of speaker-specific characteristics by employing feature vectors extracted from the speech signal. Subsequently, two main categories of features arise, viz. physiological and psychological or low-level and high-level.

The vast majority of speaker recognition systems are based primarily on using low-level spectral features that model a person’s vocal tract shape via Gaussian Mixture Models (GMMs) [5]. Generally these systems, especially state-of-the-art, rely on the mel-frequency cepstral coefficient (MFCC) feature extraction technique [6]. Such systems perform very good under clean conditions and acceptable under noisy matched conditions. Under mismatched conditions (channel, handset, ambient noise, etc.), however, performance significantly deteriorates [7]. One of the principal reasons for poor performance in these conditions is because of the nature of low-level features; being spectral, they are susceptible to spectral variations due to noise and channel effects [4].

Recent studies have shown that by incorporating high-level features of speech into the conventional system, the performance is improved [2], [3], [4], [8], [9], [10]. This also makes sense practically when considering the way humans use such patterns to recognize speakers, e.g. identifying impersonations.

This paper aims to investigate and further verify the impact of using a single high-level feature in conjunction with a low-level feature with respect to recognition performance. As stated above, high-level features exist in a hierarchal structure with increasing complexity. The author opted to begin testing fusion systems that comprised of single high-level features (ranging from less to more complex) in conjunction with the popular low-level MFCC features. This translated into using a prosodic feature of speech first.

Prosodic features are known to carry speaker-specific information like melody, intonation and loudness. Melody and intonation, comprising a major segment of prosody, are parameterized by the pitch (fundamental frequency - F_0) [9]. Prosodic information can be applied in two ways. Again, one proves superior in performance at the cost of complexity (viz. time and computation). The first (and quickest approach) uses the global statistics of a prosodic feature [4] that can be extracted from the same short segment of speech that the low-level MFCC feature is extracted from. The other way that

prosodic information is used is by capturing temporal dynamic changes of the speech's prosodic sequence. This information is obtained from much larger segments of speech than that of the global statistic approach.

Wildermoth and Paliwal [11] presented a technique called the Maximum Auto-Correlation Values (MACV) that extracted the global statistics of pitch and voicing features (prosodic) from the speech signal. They investigated the feature vector in a speaker identification environment using the TIMIT, NTIMIT and IISC databases. Their fusion-system, MACV-LPCC (Linear Predictive Coding Coefficients) consistently yielded improved recognition performance over the stand-alone low-level LPCC system.

Sanderson and Paliwal [10], [12] extended the application of this feature-fusion technique to a speaker verification system, concatenating the MFCC vector with MACV and obtained similar improved performances.

The work covered in this paper uses such a fusion system as in [12] but in a speaker identification environment. The work here is integral to the broader study of investigating all higher-level features. The system of the experiments investigates the lowest and simplest end of high-level features in conjunction with the best of low-level features, viz. MFCC. Consequently, improvements in recognition performance with such systems give rise to the promise of increased performances with more complex high-level features [4].

II. OVERVIEW OF SPEAKER RECOGNITION

Speaker recognition is comprised of two main areas, viz. speaker verification (SV) and speaker identification (SID). Speaker identification is concerned with recognizing an individual from a group of speakers based on a sample of his/her speech, whereas speaker verification is concerned with verifying that an individual is who he/she claims to be [13].

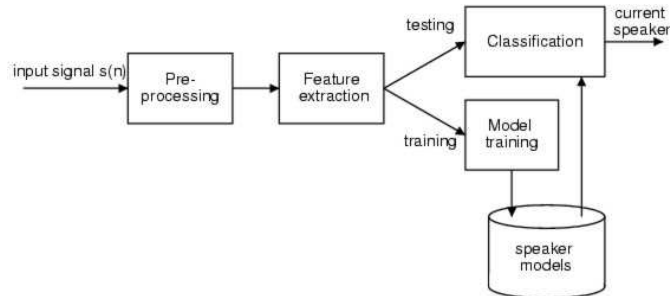


Fig. 1. Speaker recognition process overview (after [19])

Figure 1 depicts a typical speaker recognition system. As illustrated, it usually consists of four main components:

- 1) Pre-processing: speech data is prepared before the actual recognition process is performed and includes tasks like filtering, segmenting the speech into predefined time frames, etc.
- 2) Feature extraction: responsible for reducing the amount of data required to represent the input speech signal and minimising sources of noise. It does this by extracting speech vectors that preserve distinguishing speaker-specific information

- 3) Model training: responsible for creating a model of each speaker's input speech from features extracted during training phase.
- 4) Classification: During testing the models created in the training phase is made available to be tested against.

With regard to the recognition process, verification and identification are identical except for the classification component. Identification performs a $1 : N$ classification, where N is the number of speakers enrolled in the system. Verification works with a special classification case of $N = 2$, the claimant and a background model of all the speakers. The decision will tilt in one of two directions. Hence the classifier will either accept/reject a claim (verification) or return who the most likely speaker is for an unidentified speaker's utterance (identification).

The primary scope of this writing with regard to the recognition discipline is that of a text-independent speaker identification environment. This type of speaker identification is concerned with determining who, from a group of known speakers, is speaking, regardless of what is being spoken [14].

The high-level prosodic feature, MACV, to be used in the fusion system would fall under the feature extraction block and the algorithm to generate the vector follows.

III. THE MACV FEATURE VECTOR

Given a speech frame $\{s(n), n = 0, 1, \dots, N_s - 1\}$, the MACV features are computed as follows [6-8]:

- 1) Compute the autocorrelation function:

$$R(k) = \frac{1}{N} \sum_{n=0}^{N_s-1-k} s(n)s(n+k) \quad k = 0, \dots, N_s-1 \quad (1)$$

- 2) Normalise $R(k)$ by its maximum value i.e.

$$\hat{R}(k) = \frac{R(k)}{R(0)} \quad (2)$$

- 3) Divide the higher portion of $\hat{R}(k)$ into M equal parts.
- 4) Find the maximum value of $\hat{R}(k)$ in each of the M divisions.
- 5) The M Maximum Autocorrelation Values (MACV) forms an M -dimensional feature vector.

Figure 2 conceptualises the above algorithm.

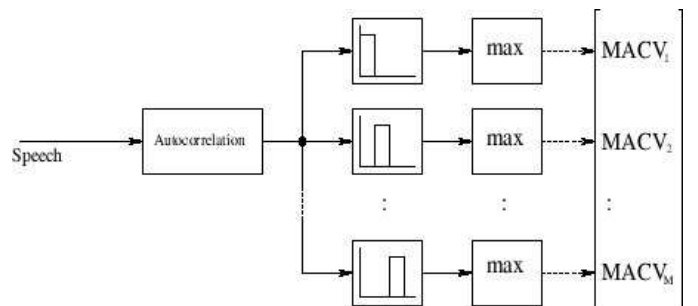


Fig. 2. MACV feature extractor (after [10])

It should be noted that the lower portion of the normalised auto-correlation function is not used. It contains information

from the vocal tract (system component of speech) which is already extracted by the MFCC vector to which the MACV will be concatenated. The higher portion of the normalised auto-correlation function was based on the fact that human pitch frequency is typically between 60-400Hz (males: 60-160Hz; females: 160-400Hz) which translates into a range from 2ms to 16ms [11].

IV. THE SPEECH DATABASE

All experiments in this research use the NTIMIT speech database. This database contains phonetically rich speech that was captured in a sound booth. The speech was then transmitted through a carbon-button telephone handset and recorded over local and long distance telephone loops. The type of noise contaminating the speech database is thus mainly caused by telephone transmission effects [15], [16]. The NTIMIT database consists of 630 speakers each having spoken 10 utterances of about 3 seconds each. The first two utterances, labeled as the sa# utterances, are common across all speakers. The next eight are all different and are labeled as the si# and sx# utterances. As a result, this database allows one to evaluate the performance of a text-independent speaker identification system using short testing and training times on telephone-quality speech.

Over the years, the NTIMIT database has been used extensively in speaker recognition tasks [15], [17]. Recently, however, researchers have criticised the NTIMIT database since the speech samples that it contains are actually read sentences which have been recorded in a single session [18]. As a result, effects caused by intersession, handset microphones and conversational speech cannot be examined with this database. Since the work presented in this research is in its early stages and is meant to primarily determine the impact of a low-end high-level feature, the database was deemed adequate in assessing the applicability of MACVs to speaker ID.

V. EXPERIMENTAL EVALUATION

This section describes the baseline system (no fusion) and then presents the results of the fusion system using the concatenated M -dimensional MACV and MFCC feature vectors extracted from the same portion of speech (short-term).

A. The Baseline System

MFCCs extracted from the input speech signal, were generated as follows.

The incoming speech signal was first multiplied by an overlapping Hamming window which divided it into a sequence of 20ms frames with an overlap of 10ms between frames. These speech frames were then Fourier transformed into the frequency domain where a sequence of log-magnitude spectra were computed. To obtain the mel-frequency cepstral coefficients, these log-magnitude spectra were filtered by a bank of mel-scaled triangular filters distributed over a bandwidth of 0Hz to 3800Hz. The outputs of the filter bank were then discrete cosine transformed into multi-dimensional MFCC feature vectors. The MACVs were generated using

the algorithm described in section III and appended to these MFCC vectors.

In order to model the distribution of feature vectors obtained for each speaker, Gaussian mixture models (GMM) were used [5], [6]. A GMM can be viewed as a non-parametric, multivariate PDF model that is capable of modelling arbitrary distributions and is currently the most dominant method of modelling speakers in speaker recognition research. The GMM of the distribution of feature vectors for speaker S is a weighted linear combination of M unimodal Gaussian densities $b_i^s(x)$, each parameterized by a mean vector μ_i^s and a covariance matrix Σ_i^s . These parameters are collectively represented by the notation

$$\lambda_s = \{p_i^s, \mu_i^s\} \quad \text{for } i = 1, 2, \dots, M \quad (3)$$

where p_i^s are the mixture weights satisfying the constraint

$$\sum_{i=1}^M p_i^s = 1 \quad (4)$$

For a feature vector x , the mixture density for speaker S is computed as

$$p(x | \lambda_s) = \sum_{i=1}^M p_i^s b_i^s(x) \quad (5)$$

where

$$b_i^s(x) = \frac{1}{(2\pi)^{\frac{D}{2}} |\Sigma_i^s|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(x - \mu_i^s)' \Sigma_i^s (x - \mu_i^s)\right) \quad (6)$$

Given a sequence of feature vectors $X = \{x_1, x_2, \dots, x_T\}$, which are assumed to be independent, the log-likelihood of a speaker model λ_s is given by

$$L_s(X) = \log p(X | \lambda_s) = \frac{1}{T} \sum_{t=1}^T \log p(x_t | \lambda_s) \quad (7)$$

For speaker identification, equation (7) is computed for the model of each speaker enrolled in the system. The identity of the speaker associated with the highest scoring model is then returned as the identified speaker. In this work GMMs with 32 mixtures to model each speaker were utilised.

B. Experimental Results

This section presents the performance of the prosodic-spectral fusion system. Note that all experiment results were averaged over three runs.

The NTIMIT database consisting of 168 speakers (112 male and 56 female) were used in the experiments. The first eight alpha-numerically numbered sentences of each speaker were utilised to train the GMMs and the last two sentences were used to test the system. In Figure 3 it can be seen that by simply appending 5 MACVs to MFCC feature vectors with varying dimensions the speaker identification performance improves in all cases. This figure also shows that a 20-dimensional MFCC feature vector results in the highest identification rate both with and without the addition of MACVs. However, the addition of MACVs improved the best identification rate from 82.74% to 85.32% - a relative improvement of over 3%.

5 MACVs was initially chosen as this was the amount of MACVs used by Wildermoth and Paliwal [11]. In order to determine whether 5 MACVs is indeed the optimal number of

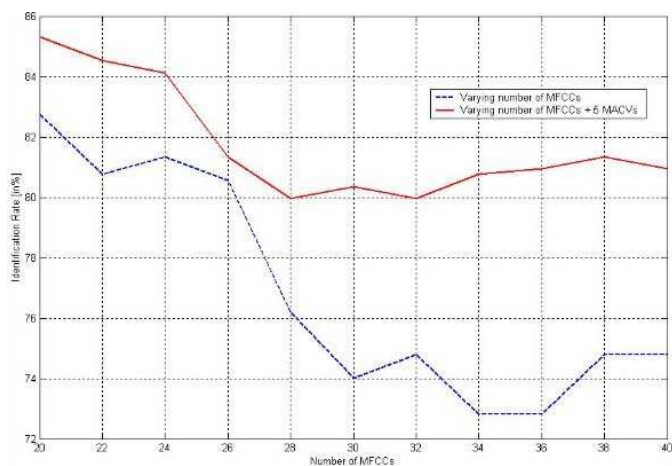


Fig. 3. SID rate versus varying number of MFCCs (constant 5 MACVs)

MACVs to use, we varied the number of MACVs appended to the 20-dimensional MFCC feature vector between 0 and 10. Our results in Figure 4 show that increasing the number of MACVs from 0 to 5 leads to a consistent improvement in performance. However, increasing the number of MACVs beyond 5 degrades system perform. This observation confirms that 5 MACVs leads to the best performance when combined with MFCCs. At this stage, however, it is unclear why this trend in performance exists.

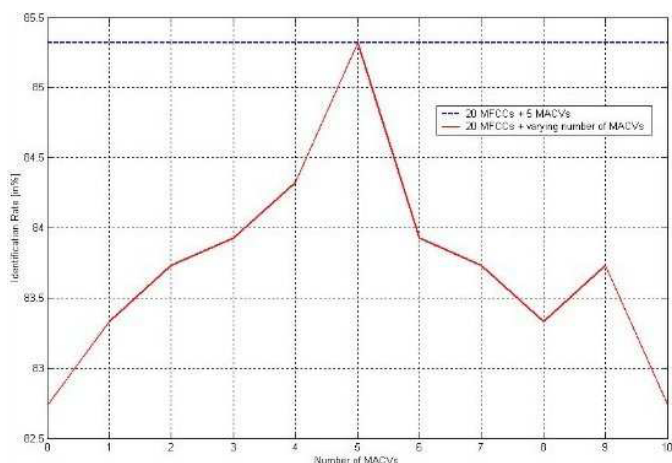


Fig. 4. SID rate versus varying number of MACVs (constant 20 MFCCs)

VI. CONCLUSION

The primary objective of this paper was to investigate the performance of a combination of a simple low-end high-level feature, viz. MACV, and a state-of-the-art low-level spectral feature, viz. MFCC, opposed to using only the spectral feature.

In this work, a relative improvement of over 3% was observed in the identification rate when a 20-MFCC vector was concatenated with 5 MACVs. This is a performance boost on using the 20-MFCC vector alone.

The results of this paper supports existing literature that says that the combination of physiological and psychological

features improve speaker recognition, viz. speaker ID over a telephone network (in this case). The increase in performance on a popular, state-of-the-art feature vector system, like the MFCC, creates anticipation for promising results to future work on other similar fusion systems performed on more challenging databases incorporating more complex high-level features. The fact that a simple, low-end prosodic feature improved the recognition performance adds to the promise of higher recognition rates in systems incorporating better high-level low-level feature fusion.

This also serves as a step forward in solving the digital divide between the lay man and information. Better speaker recognition systems would lead to more information being made accessible via public medium e.g. telephones.

REFERENCES

- [1] J.P. Campbell, Jr., "Speaker Recognition: A Tutorial," Proceedings of the IEEE, vol. 85, no. 9, pp. 1437-62, Sept. 1997.
- [2] S. Kajari, L. Ferrer, A. Ventkataraman, K. Sonmez, E Shriberg, A. Stolcke, H. Bratt, and R.R. Gadda, "Speaker recognition using prosodic and lexical features," Workshop on Automatic Speech Recognition and Understanding, St. Thomas, VI, Nov. 2003.
- [3] F. Farahani, P.G. Georgiou, and S.S. Narayanan, "Speaker identification using supra-segmental pitch pattern dynamics," ICASSP '04, Montreal, Que., Canada, May 2004.
- [4] D.A. Reynolds, W. Andrews, J. Campbell, J. Navratil, B. Peskin, A. Adami, Q. Jin, D. Klusacek, J. Abramson, R. Mihaescu, J. Godfrey, D. Jones, and B. Xiang, "The SuperSID Project: Exploiting high-level information for high-accuracy speaker recognition," ICASSP '03, Hong Kong, China, April 2003.
- [5] D.A. Reynolds, T.F. Quatieri, and R.B. Dunn, "Speaker verification using adapted Gaussian speaker mixture models," Digital Signal Processing, vol. 10, pp. 181-202, 2000.
- [6] D.A. Reynolds, and R.C. Rose, "Robust Text-Independent speaker identification using Gaussian Mixture speaker models," IEEE Transactions on Speech and Audio Processing, vol. 3, no. 1, pp. 72-83, Jan. 1995.
- [7] S. Krishnakumar, K.R. Prasanna Kumar, and N. Balakrishnan, "Pitch maxima for robust speaker recognition," ICASSP '03, Hong Kong, China, April 2003.
- [8] A.G. Adami, R. Mihaescu, D.A. Reynolds, and J.J. Godfrey, "Modeling prosodic dynamics for speaker recognition," ICASSP '03, Hong Kong, China, April 2003.
- [9] H. Ezzaïdi, and R. Jean, "Pitch and MFCC dependant GMM models for speaker identification systems," Canadian Conference on Electrical and Computer Engineering, Niagra Falls, Ont., Canada, May 2004.
- [10] C. Sanderson, and K.K. Paliwal, "Joint cohort normalization in a multi-feature speaker verification system," Proc of 10th Annual IEEE International conference on Fuzzy System, Melbourne, Vic., Australia, Dec. 2001.
- [11] B. Wildermoth, and K.K. Paliwal, "Use of voicing and pitch information for speaker recognition," Proc. 8th Australian Int. Conf. Speech Science and Technology, Canberra, 2000.
- [12] C. Sanderson, and K.K. Paliwal, "Information fusion for robust speaker verification," in Proc. Eurospeech '01, Scandinavia, 2001.
- [13] D.A. Reynolds, "A Gaussian Mixture Modeling Approach to Text-Independent Speaker Identification," Ph.D. Thesis, Georgia Institute of Technology, September, 1992.
- [14] D.A. Reynolds, "An overview of automatic speaker recognition technology," ICASSP '02, Orlando, Florida, May 2002.
- [15] H. Gish, and M. Schmidt, "Text-Independent Speaker Identification," Proc. of IEEE Signal Processing Magazine, 1994.
- [16] D.A. Reynolds, "Large Population Speaker Identification Using Clean and Telephone Speech," IEEE Signal Processing Letters, 1995.
- [17] P.J. Moreno, "Speech recognition in Telephone Environments," MSc dissertation, 1992.
- [18] D.J. Mashao, "Auditory-based speaker identification system," PRASA '01, 2001.
- [19] Tampere University of Technology Digital Media Institute: Audio Research Group. Speaker Recognition. Available: <http://www.cs.tut.fi/sgn/arg/kujahalk/speaker/>