

# Using high-level and low-level feature concatenation for speaker identification

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

Department of Electrical Engineering  
University of Cape Town, Rondebosch, 7700, South Africa

[bappanna@crq.ee.uct.ac.za](mailto:bappanna@crq.ee.uct.ac.za) [mskosan@crq.ee.uct.ac.za](mailto:mskosan@crq.ee.uct.ac.za) [daniel@eng.uct.ac.za](mailto:daniel@eng.uct.ac.za)

## Abstract

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) features like conversational, lexical, phonemic and prosodic patterns found in speech.

In this work we investigated the performance of a speaker ID system evaluated on the NTIMIT database employing the popular MFCC feature vector concatenated with a high-level feature vector containing prosodic information, viz. voicing and pitch. The vector contains the maximum autocorrelation values of a segmented frame of speech and is accordingly named the MACV feature. This paper is an extension of the work done by Wildermoth and Paliwal [11] who reported on an improved speaker ID system that used a fused LPCC-MACV feature set instead of a LPCC-only system.

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. This result corroborated with Wildermoth and Paliwal's [11] performance (an increase from 78.4% to 86.8%) and supports literature on improved recognition systems due to high-level low-level feature fusion. The increase in performance on a popular, state-of-the-art feature vector, like the MFCC, further creates anticipation for promising results to future work on similar systems used on more challenging databases.

## 1. 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]. On the other hand, the latter speaker-specific characteristics are psychological or habitual rather than physiological. They include features like conversational, lexical, phonemic and prosodic patterns found in speech [3]. 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 [4].

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, 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 [2-4], [8-10]. This also makes sense practically when considering the way humans use such patterns to recognize speakers, e.g. identifying impersonations.

Prosodic features are among the most common high-level feature used in such fusion-system research [8]. Prosodic features are known to carry speaker-specific information like melody, intonation and loudness. They are sometimes referred to as source features because they originate at the glottal source [11]. Melody and intonation, comprising a major segment of prosody, are parameterized by the pitch (fundamental frequency – F0) [9]. Many past efforts using stand-alone pitch features returned unimpressive results. The main reason for such performance was attributed to poor, unreliable pitch estimation methods [11].

Wildermoth and Paliwal [11] presented a technique called the Maximum Auto-Correlation Values (MACV) that extracted pitch and voicing information from the speech signal. They investigated the feature vector in a speaker identification environment using the TIMIT, NTIMIT and IISC databases. The performance of the SID system using only a MACV feature vector was poor, although it was an improvement on systems using conventional pitch features only. Results were greatly improved, though, when MACV was combined with a cepstral feature vector, viz. the LPCC feature vector. On all databases there was a significant improvement with the fusion system than with LPCC alone. Experiments on the NTIMIT database, for example, yielded an ID rate improvement of 78.4% to 86.8% (a relative improvement of 10.71%).

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.

In this paper, we further explore the work done in [11] by also working in an identification scenario but with the exception of concatenating the more common MFCC feature vector with MACV. Experiments are investigated on the NTIMIT database.

We use the same concatenation technique used in [11] with the aim of verifying its superior performance to the MFCC extraction technique alone. The primary objective is to corroborate the increase in robustness and performance of a speaker identification system by combining current, good performing low-level features with high-level features.

## 2. Overview of speaker identification

Speaker identification, together with speaker verification, makes up the larger discipline of speaker recognition. 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].

In this paper, research is conducted into the area of text-independent speaker identification. 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]. In literature this is referred to as closed-set identification as the system must perform a 1:  $N$  classification, where  $N$  is the number of speakers enrolled in the system [13]. The speaker identification process can be summarised as follows.

First the system needs to be trained with samples of speech collected from the speakers to be identified. Once this is complete, the system is tested (a speaker is identified) by comparing a speech sample from an unidentified speaker to the speech samples stored by the system and determining who the most likely speaker is [14].

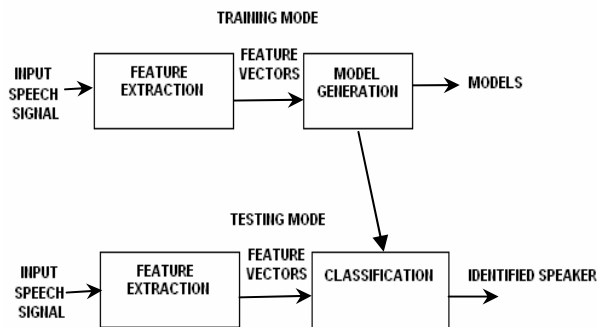


Figure 1: A typical speaker identification system

Figure 1 depicts a typical speaker identification system. As illustrated, it usually consists of three main components. These components perform the following tasks:

The *feature extraction component* is responsible for reducing the amount of data required to represent the input speech signal and minimising sources of noise. It does this while hopefully preserving distinguishing speaker-specific information. The *model generation component* is responsible for creating a model of each speaker's speech characteristics during training. During testing it makes its database of speaker models available to the classification component. The *classification component* is used during testing to compare an unidentified speaker's utterance to

the speaker models produced by the model generation component. On the basis of these comparisons, it determines who the most likely speaker is.

The MACV feature set would fall under the feature extraction block and the algorithm to generate the vector follows.

## 3. 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]:

- a) Compute the autocorrelation function:

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

- b) Normalize  $R(k)$  by its maximum value i.e.

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

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

Figure 2 conceptualises the above algorithm.

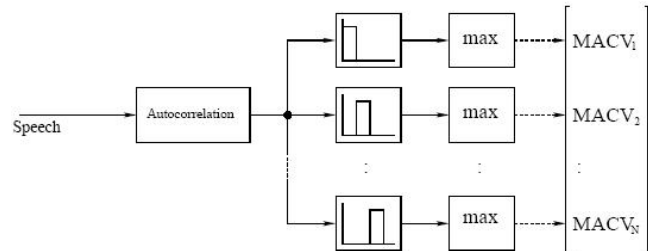


Figure 2: MACV feature extractor (after [10])

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].

## 4. 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 verify observations made by other researchers, the database was deemed adequate in assessing the applicability of MACVs to speaker ID.

## 5. Experimental Evaluation

In this section results are presented concerning the concatenation of the  $M$ -dimensional MACV feature vector with the MFCC feature vector extracted from the same portion of speech.

### 5.1. The Baseline System

In this work the feature extraction component extracts MFCCs as well as MACVs from the input speech signal. The MFCCs 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 feature vectors. The MACVs were generated using the algorithm described in *section 3*.

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  $b_i^s(\mathbf{x})$ , each parameterized by a mean vector  $\boldsymbol{\mu}_i^s$  and a covariance matrix  $\Sigma_i^s$ . These parameters are collectively represented by the notation

$$\lambda_s = \{p_i^s, \boldsymbol{\mu}_i^s, \Sigma_i^s\} \quad \text{for } i = 1, \dots, M \quad (3)$$

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  $\mathbf{x}$ , the mixture density for speaker  $S$  is computed as

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

where

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

Given a sequence of feature vectors  $X = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_T\}$ , which are assumed to be independent, the log-likelihood of a speaker model  $\lambda_s$  is given by

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

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.

### 5.2 Experimental Results

Experiments were primarily performed to verify that appending MFCC feature vectors with MACV features does indeed improve speaker identification performance. Note that all experiment results were averaged over three runs.

For our experiments we used the ‘test’ portion of the NTIMIT database consisting of 168 speakers (112 male and 56 female). We used the first eight alpha-numerically numbered sentences of each speaker to train the GMMs and the last two sentences were used to test the system. In Figure 3 we show that by simply appending 5 MACVs to MFCC feature vectors with varying dimensions improves speaker identification performance 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 identification rate from 82.74% to 85.32% - a relative improvement of over 3%.

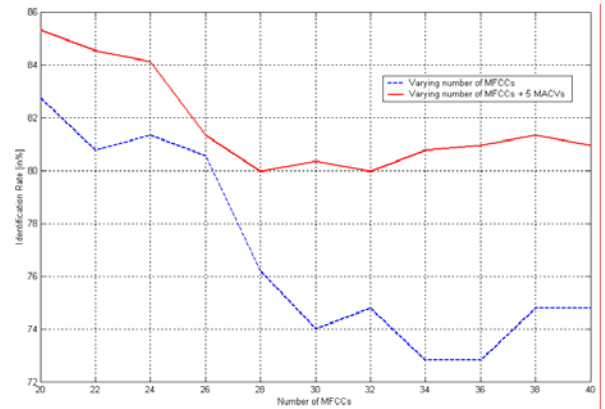


Figure 3: Speaker identification rate versus varying numbers of MFCCs (with and without the addition of 5 MACVs)

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 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 performance. 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.

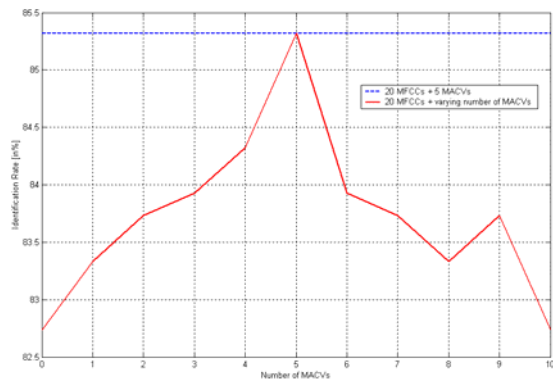


Figure 4: Speaker identification rate versus varying numbers of MACVs (with the addition of 20 MFCCs)

## 6. Conclusion

The primary objective of this paper was to investigate the performance of a combination of a high-level feature, viz. MACV, and a popular low-level cepstral feature opposed to using only the cepstral feature. In doing so, a secondary purpose arose to extend the work done by Wildermoth and Paliwal [11] who combined MACV and LPCC features in a SID environment; the fusion of MACV and the popular MFCC feature was investigated in this paper in a SID scope on the NTIMIT database.

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 an improvement on using the 20-MFCC vector alone.

Compared to the work by Wildermoth and Paliwal [11] who used MACV and LPCC, the 3% relative improvement in ID rate was less than their 10%. It is worthy to note, though, that their LPCC-alone system yielded an ID rate of 78.4% whereas the MFCC-alone system investigated in this paper yielded an 82.74% ID rate. So, although the relative improvement rates were different, overall recognition performance was about the same.

The results of this paper supports existing literature that says that the combination of physiological and psychological features improve speaker recognition, specifically speaker ID in this case. 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 other similar fusion systems performed on more challenging databases.

## 7. 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. Kajarekar, L. Ferrer, A. Ventkataraman, K. Sonmez, E. Shriberg, A. Stolcke, H. Bratt, and R.R. Gadde, "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. Ezzaidi, 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 10<sup>th</sup> 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. 8<sup>th</sup> 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.