Features for Phoneme Independent Speaker Identification

Jianglin Wang, An Ji, Michael T. Johnson
Speech and Signal Processing Laboratory
Department of Electrical and Computer Engineering
Marquette University, Milwaukee, USA
{jianglin.wang, an.ji, mike.johnson}@marquette.edu

Abstract

This paper describes a unique cross-phoneme speaker identification experiment, using deliberately mismatched phoneme sets for training and testing. The underlying goal is to identify features that represent broad individually unique characteristics rather than those that represent phonetic differences, as are more typical of modern speaker identification and verification systems. A wide range of features are proposed and evaluated within this context using a Gaussian Mixture Model framework. The results show that log-area ratio has better phonetic independence than MFCCs, that residual phase carries substantial speaker information, and identifies several other features that also have usefulness for speaker identification independent of phonetic content.

1. Introduction

The task of speaker identification and verification has received a great deal of attention from the research community in the past decade, and there have been substantial gains in accuracy as well as channel and background robustness [1, 2]. However, the fundamental mechanism of state-of-the-art systems has remained phonetic rather than physiological in nature, and little progress has been made toward identifying individually unique speech characteristics that are independent of phonetic content.

The most common approach for current identification systems is based on Gaussian Mixture Models (GMM) [3] or GMMs coupled with Support Vector Machines (GMM-SVM) [4, 5]. Verification is accomplished through likelihood comparison with appropriate cohort models, as with a Universal Background Model (UBM) system [6]. Standard spectral features such as Mel Frequency Cepstral Coefficients (MFCCs) with energy and first and second order derivatives are used, typically projected to a lower dimension through Heteroscedastic Linear Discriminant Analysis (HLDA) [7]. A variety of feature transformation techniques are used, including Nuisance Attribute Projection and Latent Variable Analysis [8, 9]. Score normalization techniques are also incorporated, such as HNorm, TNorm, and others. In the past few years it has become commonplace, for example in the NIST Speaker Identification challenges, to build a large number of different system types and integrate them through various forms of intelligent score fusion. In all cases, though, the basic features remain primarily spectral in nature, relating to vocal tract parameters that are heavily correlated with phonetics.

There is still a significant need for identification of unique speaker-specific features that are more independent of phonetic content and representative of more underlying physiological characteristics. Examples of such features might be those more related to source excitation rather than the vocal tract, such as derivatives of fundamental frequency characteristics like jitter, shimmer, and harmonic amplitudes [10]. Orthogonal linear prediction coefficients have been proposed as features for speaker identification [11] because they are more constant across utterances and thus are more independent of the linguistic independence and indicative of the speaker. Recently, it has been reported that LP residual phase also contains speaker-specific source information [12].

This paper uses a unique cross-phonetic experimental design in order to investigate and evaluate a number of features, some previously proposed and some newly identified. The experimental design incorporates phonetically separate training and testing sets of vowels, chosen on the basis of separation in the F1/F2 vowel triangle space. The goal is to focus primarily on phonetically independent characteristics rather than those that maximize accuracy on traditional speaker identification tasks.

978-1-4673-0174-9/12/$31.00 ©2012 IEEE

This paper is organized as follows. Section 2 provides the details of each feature extraction method. The experiment data, classification method and results are described in Section 3. Final discussions and conclusions are given in Section 4.

2. Feature sets

The proposed speaker-specific features include log area ratio (LAR), modified residual phase, fundamental frequency, jitter, shimmer, shimmer of the LAR, shimmer of the LPC coefficients, average harmonic amplitude difference, mean residual phase, mean fundamental frequency, and variance of residual phase. Of these, the modified residual phase, average harmonic ratio difference, LAR shimmer, and LPC shimmer have not been previously proposed for speaker identification. The baseline features used for comparison are MFCCs, consisting of 12 MFCCs plus energy and delta/delta-delta MFCCs forming a 39 dimensional feature vector.

2.1. Log area ratio

Log area ratio coefficients are well-known spectral measures typically derived from the linear prediction coefficients 0, modeled as reflection coefficients of a non-uniform acoustic tube vocal tract. The definition is:

$$LAR_i = \log\left(\frac{A_i}{A_{i+1}}\right) = \log\left(\frac{1+\alpha_i}{1-\alpha_i}\right), \ \mathrm{A}_{p+1}=1 \qquad (1)$$

where α_i represent the parcor coefficients. If we denote $a_i^{(k)}$ as the i^{th} LPC for a k^{th} pole linear prediction model, then $\alpha_i = a_i^{(i)}$, i=1,⋯,p.

2.2. Residual phase

The original definition of residual phase is the cosine of the phase function of the analytic signal 0. The analytic signal is derived from the LP residual of a speech signal. In this experiment, the RP coefficients are modified by including the sine of the phase information in addition to the cosine. The calculation of RP is shown below.

$$r(n) = s(n) - \hat{s}(n) = s(n) + \sum_{k=1}^{p} a_k s(n-k) \qquad (2)$$

where p is the order of prediction and { a_k } are the linear prediction coefficients obtained from LPC analysis.

The analytical signal $r_a(n)$ is given by

$$r_a(n) = r(n) + jr_h(n) \qquad (3)$$

where $r_h(n)$ is the Hibert transform of r(n). The cosine of the phase information is calculated by the following equation:

$$\cos\left(\theta(n)\right) = \frac{R_e(r_a(n))}{\left|r_a(n)\right|} \qquad (4)$$

The sine of the phase information is

$$\sin\left(\theta(n)\right) = \sqrt{1-\cos^2\left(\theta(n)\right)} \qquad (5)$$

The new modified RP coefficients represent co-ordinates on a unit circle,

$$RP = \left\{\cos\theta, \sin\theta\right\} \qquad (6)$$

2.3. Fundamental frequency

Fundamental frequency is extracted from the phoneme utterance of a speaker using the COLEA toolbox 0 cepstrum implementation. Results are post-processed by median filtering. Unvoiced frames are considered to have a frequency of zero.

2.4. Jitter and shimmer

Jitter is the relative evaluation of the period-to-period variability of the pitch within a frame 0, defined as:

$$\mathrm{Jitt} = \frac{\left|To^{(i)} - To^{(i+1)}\right|}{\frac{1}{\mathrm{N}}\sum_{i=1}^{N} To^{(i)}} \qquad (7)$$

where $To^{(i)}$, $i=1,\ 2\ ...\ N$ is the extracted pitch period of the i^{th} frame and N is the total number of voiced frames in the utterance.

Shimmer is relative evaluation of the period-to-period variability of the peak-to-peak amplitude,

$$\mathrm{Shim} = \frac{\left|A^{(i)} - A^{(i+1)}\right|}{\frac{1}{\mathrm{N}}\sum_{i=1}^{N} A^{(i)}} \qquad (8)$$

where $A^{(i)}$, $i=1,\ 2...N$ is the extracted peak-to-peak amplitude and N is the total number of voiced frames in the utterance.

2.5. Shimmer of LAR and LPC

The definition of shimmer defined in equation (8) is modified to calculate the shimmer of LAR and LPC. The shimmer of LAR measures fluctuations of the peak-to-peak value of log area ratio within each frame.

$$\mathrm{ShimLAR} = \frac{\left|LAR^{(i)} - LAR^{(i+1)}\right|}{\frac{1}{\mathrm{N}}\sum_{i=1}^{N} LAR^{(i)}} \qquad (9)$$

where $LAR^{(i)}$ is the log area ratio in i^{th} frame.

The shimmer of LPC measures fluctuations of the peak-to-peak value of linear prediction coefficients within each frame.

$$\text{ShimLPC} = \frac{\left|LPC^{(i)} - LPC^{(i+1)}\right|}{\frac{1}{\text{N}}\sum_{i=1}^{N} LPC^{(i)}} \tag{10}$$

where $\text{LPC}^{(i)}$ is the log area ratio in i^{th} frame.

2.6. Average harmonic amplitude difference

The average harmonic amplitude difference is defined here as a function of the amplitudes of the harmonics of the fundamental frequency of a given frame:

$$AHAD = \frac{\sum_{i=1}^{N}\left|HA_i - HA_{i+1}\right|}{N-1} \tag{11}$$

where HA_i, i=1, 2, …, 6, are the first six harmonic amplitude values of the fundamental.

2.7. Statistical features

A number of statistics of the above features are also computed and used as features, including the following:

- mean of harmonic amplitude
- mean of residual phase
- variance of residual phase

2.8. Overview of feature characteristics

The proposed features in this experiment include 11 different features. Each feature reflects a different representation of a speaker's vocal characteristics. Log area ratio reflects the variation of a speaker's vocal tract and has a linear spectral sensitivity 0. The residual phase represents the excitation source characteristics of a speaker. Jitter and shimmer reflect the micro-instability of vocal fold vibrations of the speaker. The shimmer of LAR and LPC show fluctuation patterns within the spectral characteristics, and are therefore related to vocal tract stability. Average harmonic amplitude difference provides a quantitative measure of glottal abduction-adduction [15, 16]. The statistical measures similarly reflect fluctuation patterns of the various spectral and excitation metrics.

3. Experiment

In order to support the goal of evaluating phonetically-independent features for speaker identification, a cross-phonetic experimental paradigm has been designed. Using the well-known vowel triangle in the F1/F2 feature space, vowel sets with minimal spectral similarity were selected for training and testing sets. The data used is extracted from the TIMIT database. The phoneme set used in these experiments is divided primarily on the basis of the overall vowel height (i.e. primarily correlated with formant F1). For training low vowels {/ae/, /aa/, /ah/, /eh/, /ao/} are used, while the phoneme set for testing includes the high vowels {/iy/, /ih/, /er/, /ow/, /uh/, /uw/}. A subset of 25 speakers within the same dialect region is used for evaluation, leading to an overall data set consisting of 1621 phoneme utterances. Phonemes with fewer than 700 samples are discarded in order to ensure accuracy of fundamental frequency related and statistical features.

Thirty two mixture GMM are used for classification models 0. The programming toolkit HTK 3.4.1 from Cambridge University is used for all training and testing 0.

3.1. Experimental results

Each individual feature was evaluated for identification accuracy, in comparison to a baseline using a standard 12-element MFCC feature vector. Various increasing combinations of feature sets were then implemented to assess how the features could be used in combination.

Table 1 shows the results of each individual feature, while Table 2 shows the results of different feature combinations. Individually, the best overall feature is the LAR feature at 41.8% accuracy, followed by MFCCs. The highest non-spectral feature is the residual phase, with 24.0% accuracy.

The highest total accuracy of 51.9% is obtained by a combination of all the proposed features, with a feature dimension of 304. This compares to a 35.4% accuracy obtained using the baseline MFCC features.

It is interesting to observe that, despite the fact that the training and testing sets were deliberately chosen to minimize phonetic similarity, the basic spectral measures of MFCCs and LARs still outperform excitation measures based on fundamental frequency. One key observation is that LARs seem to be a substantially better spectral feature for the purposes of phonetically-independent speaker identification, and that the excitation-related measures are at least able to contribute to an overall increase in accuracy when combined with the LAR. Of the excitation-related measures, the residual phase is clearly the strongest individual component.

Table 1. The classification accuracy of individual features

Individual Feature (Dimension)	Accuracy (%)
MFCC(39)	35.4
Log Area Ratio(39)	41.8
Jitter(1)	10.0
Shimmer(1)	4.3
Residual Phase(256)	24.0
Fundamental Frequency(1)	13.9
LPC Shimmer(1)	6.5
LAR Shimmer(1)	6.0
Average harmonic amplitude difference(1)	7.3
Mean of LAR(1)	5.6
Mean and Variance of RP(2)	15.6

Table 2. Classification results for different feature combinations

Feature Combination (Dimension)	Accuracy (%)
MFCC(39)	35.4
LAR + Jitter + Shimmer + F0(42)	43.6
LAR + Jitter + Shimmer + RP + F0 + Shimmer of LPC & LAR(300)	45.8
LAR + Jitter + Shimmer + RP + F0 + Shimmer of LPC & LAR + Harmonic Amplitude(301)	48.1
LAR + Jitter + Shimmer + RP + F0 + Shimmer of LPC & LAR + Harmonic Amplitude + Statistics of RP & LAR(304)	51.9
Proposed Feature Group + MFCC(343)	50.5

4. Conclusions

A novel cross-phoneme experimental paradigm has been introduced for evaluating phonetically-independent features for speaker identification. Initial results suggest that the LAR is a better spectral metric than MFCCs, and that a number of excitation-related metrics, in particular the residual phase, are also helpful in this context.

References

[1] J.P. Campbell, “speaker recognition: a tutorial,” in *Proc. IEEE*, vol. 85, no. 9, pp. 357-366, Aug. 1980.

[2] N. Zheng, T. Lee, and P. C. Ching, “Integration of complementary acoustic features for speaker recognition,” *IEEE Signal Proc*. Letters, vol. 14, 2006.

[3] D.A. Reynolds, “Speaker identification and verification using Gaussian mixture speaker models”, *Speech Communication*, vol 17, 91 – 108, 1995.

[4] S. Fine, J. Navratil, and R.A. Gopinath, “A hybrid GMM/SVM approach to speaker identification,” *ICASSP,* vol. 1, pp.417-420, 2001.

[5] C.H. You, K.A. Lee, and H. Li, “An SVM kernel with GMM-supervector based on the Bhattacharyya distance for speaker recognition,” *IEEE Signal Process.* Lett., vol. 16(1), pp. 49-52, Jan. 2009.

[6] A. Akula, V.R. Apsingekar, and P.L.D. Leon, “Speaker Identification in Room Reverberation Using GMM-UBM”, in *DSP/SPE IEEE workshop*, 2009.

[7] N. Kumar and A. G. Andreou, “Heteroscedastic discriminant analysis and reduced rank HMMs for improved speech recognition”, *Speech Communication,* vol.26, pp. 283–97, 1998.

[8] B. Raj and P. Smaragdis, “Latent variable decomposition of spectrograms for single channel speaker separation,” in *Proc Of IEEE on Applications of Signal Processing to Audio and Acoustics,* pp. 17–20, New Paltz, NY, USA, October 2005.

[9] A. Solomonoff, W. M. Campbell, and I. Boardman, “Advances in channel compensation for SVM speaker recognition,” in *Proc. of ICASSP,* 2005.

[10] X. Li and M.T. Johnson, etc., “Stree and emotion classification using jitter and shimmer features”, in *Proc. of ICASSP,* hawai’I, USA, 2007.

[11] R.S. Marvin, “Speaker recognition using orthogonal linear prediction”, *IEEE Trans. Acoust., Speech, Signal Processing*, vol 24, pp.283-9, Aug. 1976.

[12] K.S.R. Murthy and B. Yegnanarayana, “Combining evidence from residual phase and MFCC features for speaker recognition,” *IEEE Signal Process*. Lett., vol 13(1), pp. 52-6, Jan. 2006.

[13] R.S. Marvin, “Linear prediction: a tutorial review,” *Prco. IEEE,* vol. 63, 561-579, 1975.

[14] P. Loizou, “COLEA: A MATLAB software tool for speech analysis”, *Department of Electrical Engineering, University of Texas at Dallas,* Richardson, TX, 1999.

[15] G.D. Krom, “Some spectral correlates of pathological breathy and rough voice quality for different types of vowel fragments”, *Journal of Speech and Hearing Research,* 38, 794-811, 1995.

[16] H.M. Hanson, “Glottal characteristics of female speakers: Acoustic correlates”, *Journal of the Acoustical Society of America*, 101(2), 466-481, 1997.

[17] A. P. Dempster, M. Laird, and D. B. Rubin, “Maximum likelihood from incomplete data via the EM algorithm”, *Journal of the Royal Statistical Society,* pp. 1-39, 1997.
[18] S. Young, et al., *the HTK Book* (for HTK Version 3.4.1), 2009.