A New Set of Features for Text-Independent Speaker Identification

Carol Y. Espy-Wilson, Sandeep Manocha and Srikanth Vishnubhotla

Institute for Systems Research and Dept. of Electrical & Computer Engineering,
University of Maryland, College Park, MD, USA 20742
{espy, smanocha, srikanth}@umd.edu

Abstract

The success of a speaker identification system depends largely on the set of features used to characterize speaker-specific information. In this paper, we discuss a small set of low-level acoustic parameters that capture information about the speaker's source, vocal tract size and vocal tract shape. We demonstrate that the set of eight acoustic parameters has comparable performance to the standard sets of 26 or 39 MFCCs for the speaker identification task. Gaussian Mixture Models were used for constructing speaker models.

Index Terms: speaker identification, acoustic parameters, Gaussian Mixture Models (GMM), Mel-Frequency Cepstral Coefficients (MFCC), speaker specific features

1. Introduction

The goal of speaker identification is to determine which one of a group of known speakers best matches the test speech sample. Speaker identification can be constrained to a known phrase (text-dependent) or totally unconstrained (text-independent) [1]. Success in this task depends on extracting speaker-dependent features from the speech signal that can effectively distinguish one speaker from another. Various features have been employed in the past for speaker identification, the most popular among them being the mel-frequency cepstral coefficients (MFCCs) as they carry both speech and speaker information. Although the MFCCs implicitly capture speaker-specific information, we want to explore parameters that explicitly capture this information.

The set of Acoustic Parameters (APs) that we propose are aimed at extracting speaker-specific features from the speech signal that will help distinguish one speaker from another. Our set of features consists of four formants (F1, F2, F3, F4), the amount of periodic and aperiodic energy in the speech signal, the spectral slope of the signal and the difference between the strength of the first and second harmonics. We performed text-independent speaker identification experiments using our feature set and the standard MFCCs for populations varying from 50 to 250 speakers of the same gender.

The results show that our eight parameters have a better performance for female speakers than that of the 26 MFCCs and 39 MFCCs and on the average they have comparable performance for varying population size.

Section 2 describes the theoretical motivation and the feature set that we have employed and their computation is briefly described in Section 3. The database used and the experiments that we conducted are explained in Section 4. Section 5 discusses the results obtained, and important conclusions and future work are outlined in Section 6.

2. Motivation and Set of Features

While traditional speaker identification systems rely on the vocal tract dynamics and under-emphasize the significance of the source, more recent work has shown that the addition of source information can prove to be valuable speaker-specific information [2]. The MFCCs implicitly code the vocal tract information and some source information in them, while the Acoustic Parameters (APs) attempt to explicitly arrive at this information. To capture the difference in articulatory strategies across speakers, APs were developed that measure from the speech signal the acoustic cues associated with different voice qualities (source information) and different vocal tract configurations.

2.1. Features of the source

The source information of a speaker depends on factors such as the shape and timing of the glottal pulses, whether or not the vocal folds close completely and, the tradeoff between the glottal source and supraglottal source during voiced obstruent sounds. Based on these factors, one can describe the way a speaker sounds in terms of the voice quality of the speaker. These speaker-specific characteristics determine (a) the high-frequency roll off of the speech spectrum, (b) the relative amplitudes of the very low-frequency harmonics, and (c) the harmonic and inharmonic structure of the speech waveform respectively.

Voice qualities can be divided into several broad categories such as modal, breathy, creaky, pressed voice, etc. Klatt and Klatt [3] contrast the difference in acoustic properties of speech produced with a modal voice, a breathy voice and a pressed voiced. During modal speech, the speech spectrum is harmonic throughout with a high-frequency roll off of 12 dB/octave. However, in breathy phonation, the vocal folds do not close completely and the glottal waveform is more sinusoidal. This leads to a steeper high-frequency roll off and a very prominent first harmonic. In addition, the higher frequencies, starting around F3 are inharmonic due to aspiration. In pressed speech, the vocal folds close more abruptly which leads to considerably more high frequency energy relative to that seen in modal speech. Finally, in creaky voice, the vocal folds have an irregular vibration pattern so that the fundamental frequency is usually very low.

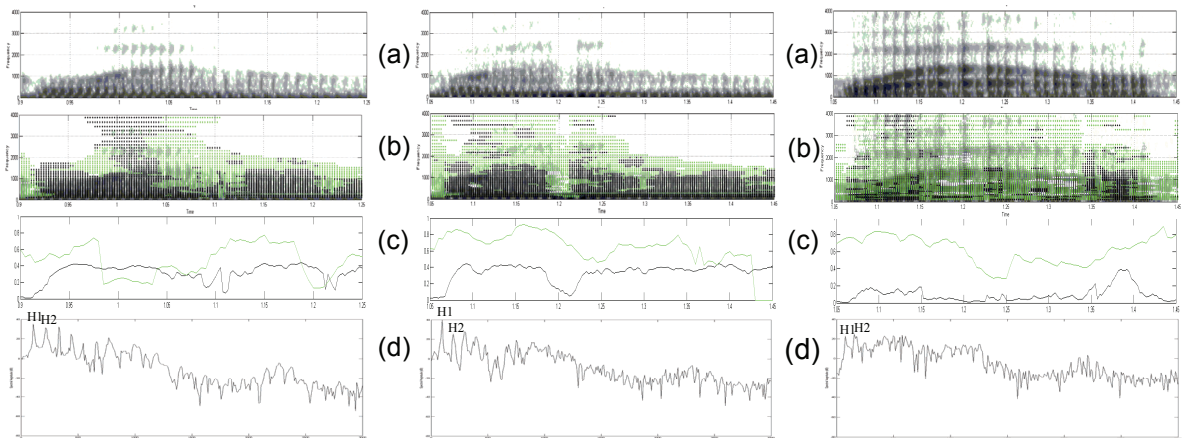

Fig 1 : Spectrogram (a) & Spectro-Temporal Profile (b) for a Modal (Left), Breathy (Middle) and Creaky voice (Right); green dots represent aperiodicity and black dots represent periodicity content. The figures (c) show the summary aperiodicity (green) and periodicity (black) measure, while (d) show a spectrum slice for each case

The parameters that we used to capture the differences in what is happening at the source are: (1) the spectral slope, (2) the difference between the amplitudes of the first and second harmonic (H1-H2), (3) the proportion of periodic energy in the speech signal and (4) the proportion of aperiodic energy in the speech signal. As an illustration, Figure 1 shows the difference in these parameters for a portion of an utterance produced by same speaker using his modal voice, a breathy voice and a creaky voice. The proportion of periodic and aperiodic energy was determined by our Aperiodicity Periodicity and Pitch (APP) Detector [4]. The APP detector generates a spectro-temporal profile of the periodic and aperiodic regions in the speech signal (part (b) of Figure 1) and also provides a summary measure of the amount of the periodic energy and aperiodic energy in the signal.(part (c) of Figure 1). The parameter H1-H2 (part (d) of Figure 1) is found to be 3.91, 14.51 and -1.94 dB for the modal, breathy and creaky voice respectively.

At present, the APP detector does not distinguish between aperiodicity due to noise and aperiodicity due to irregular vocal fold vibration. Thus, the APP detector finds considerably more aperiodicity during creaky voice than during breathy voice.

The relative amounts of periodic and aperiodic energy during voiced sounds not only tell us about voice quality, but they also tell us about the different articulatory strategies used by speakers when producing voiced obstruents which are canonically produced with a strong supraglottal source and a weaker glottal source [5]. Figure 2 shows spectrograms of the alphabet "Z" produced by three different speakers. Speaker 1 produces the /z/ with a strong glottal source and little turbulence so that the amount of periodic energy is considerably strong than the amount of aperiodic energy throughout. Speaker 2 starts by producing the /z/ with a stronger glottal source, but transitions to producing it with a stronger turbulent source (i.e., he narrows the alveolar constriction as he is producing the /z/). Speaker 3 produces the /z/ canonically, with a strong turbulent source throughout the sound. It is hoped that tradeoffs of this kind will be captured in the Gaussian Mixture Models (GMM) framework given we are using multimodal distributions. Note that some speakers, particularly when talking casually, will produce weaker voiced fricatives like /v/ and voiced stop consonants with such a weak constriction that they will contain strong periodic energy and little, if any, aperiodic energy [6,4].

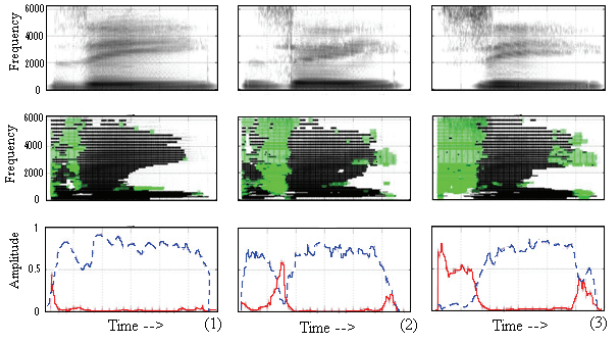

Figure 2: *Top: Spectrogram, Middle: spectro-temporal profile (periodic in black and aperiodic in green), Bottom: periodic (red solid line) and aperiodic (blue dotted line) summary information for the letter "Z" for 3 different speakers*

2.2. Features of the vocal tract

The frequencies of the formants during sonorant sounds provide information about the length and shape of the vocal tract. Generally, F1 and F2 vary considerably due to the vowel being articulated, whereas F3 and F4 change very little. The vocal tract length of a speaker can usually be characterized by F3. However, we have found that F3 and the higher formants may be indicative of vocal tract shape during sonorant consonants where narrower constrictions are produced. For example, the American English /r/ sound can be produced with a variety of vocal tract configurations that all give a very similar acoustic profile for F1-F3 [7]. However, it appears that the higher formants, in particular F4 and F5, may be acoustic signatures of the differences in vocal tract shape. Figure 3 shows spectrograms of the nonsesnse word "warav" produced by two different speakers, and flesh-point data showing tongue position during the /r/ sound. The /r/ in the word on the left is produced with the tongue dorsum high and tongue tip lower, whereas the word on the right is produced with the tongue tip high and tongue dorsum lower. The trajectories of F1, F2 and F3 between the /r/ and adjacent vowels look similar across both words. However, F4 and F5 show little movement between the /r/ and the adjacent vowels in the spectrogram on the left In contrast,

in the spectrogram on the right, F4 and F5 track F3. Thus, they are significantly lower during the /r/ sound relative to their positions during the adjacent vowels.

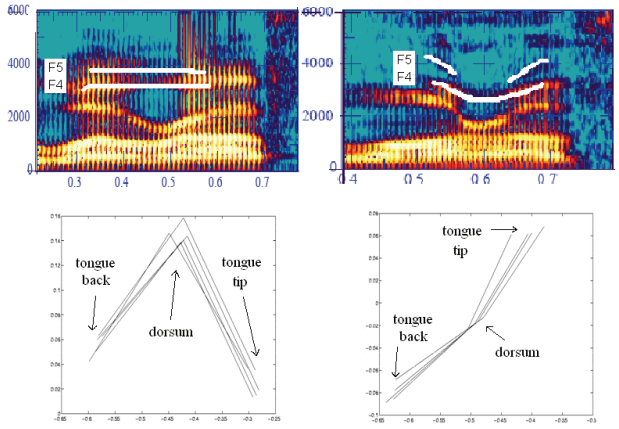

Figure 3: *Spectrogram (top) of nonsense word 'warav' said by two different speakers and flesh point data (3 palletes) showing tongue position of /r/ at the lowest point of F3.*

The parameters that we used to capture the differences in vocal tract configuration are the first four formants: F1, F2, F3 and F4. The fifth formant is not used since we are using telephone speech sampled at 8 kHz. Note that while we have thought of F4 as providing information about vocal tract shape, there is research that shows a strong relationship between F4 and the dimensions of the laryngeal cavity during vowel sounds [8, 9].

3. Automatic Computation of Features

The computation of the eight acoustic parameters described in the previous section was completely automated. The formant frequencies were computed using the ESPS Formant Tracker. The reliability of these parameters is dependent on the accuracy of the formant tracker. However we did not make any manual corrections to the formant frequency values. These errors would be partially compensated if they are systematic errors and due to the fact that we are not interested in the instantaneous formant values, but rather the range of the formant frequencies.

The proportion of periodicity and aperiodicity was calculated using the APP detector. The algorithm first computes the Average Magnitude Difference Function (AMDF) over a pre-defined lag range for all the non-silent frequency channels in a given frame. For a periodic channel, the AMDF exhibits strong minima at regular lag intervals (corresponding to the period of the signal), whereas for an aperiodic signal the AMDF has numerous minima at random lags. The locations of these minima will be coherent across the different channels for a predominantly periodic frame. The APP detector quantizes this distribution of the AMDF minima across the channels to estimate the proportion of periodicity and aperiodicity.

The spectral slope was computed by fitting a regression line to the spectrum of the signal. The slope of the line yields the spectral tilt. The first and second harmonics are located by finding the first two prominent peaks in the spectrum. Suitable thresholds were used to avoid small and insignificant peaks and two peaks that are unreasonably close to each other. The difference in the amplitudes of the harmonics (H1-H2) was used as the parameter.

4. Experiments

A number of text-independent speaker identification experiments on telephone speech were conducted. The population size of each experiment was varied from 50 to 250 speakers and they were either all male speakers or all female speakers. The speaker identification system selected the speaker from the set of speaker models that best matched the test utterance. The NIST '98 Evaluation Database was used for the experiments. Several test utterance from each speaker was used, resulting in a large number of Speaker-ID tests for both the female and male population.

4.1. Database

The NIST '98 Evaluation Database consists of telephone speech sampled at 8 kHz. The Database contains 250 male speakers and 250 female speakers. The training utterances are taken from the train/s1a/ directory and the testing utterances are taken from the test/30/ directory of the database. There is no handset variation between the training and the test utterances. The length of each training utterance is approximately 1 minute and the testing utterances are about 30 seconds in duration. An energy threshold was used to remove the silence portion (which sometimes has low amplitude background noise) from the speech. This resulted in training utterances of about 30 to 40 seconds and testing utterances of about 10 to 20 seconds.

4.2. Method

The features (both MFCCs and APs) were computed from the speech signal every 10 ms. Thirteen MFCCs were computed using cepstral mean normalization; the zeroth cepstral coefficient was not used. The set of 26 MFCCs consisted of the 13 coefficients and their derivatives. The acceleration coefficients were appended to these to obtain the 39 MFCCs.

The speaker models were constructed using the Gaussian Mixture Models (GMM) and were trained using maximum-likelihood parameter estimation [10]. The MIT-LL GMM system was used for constructing the speaker models. Various model orders were tested and we empirically determined that the 32-mixture GMM gave the best performance for the APs and the MFCCs. The test utterance was identified with the speaker whose model yields the highest likelihood for the test utterance. The accuracy of the system was computed using the identification errors made by the system. To obtain the accuracies for different population sizes, the 250 speakers of each gender were divided into groups where the number of speakers in each group is the population size. The accuracy for the particular population size is the average of the accuracies over all the groups.

5. Results

The speaker identification experiments were conducted with different feature sets. The error rates of the systems are

summarized in Table 1.

Table 1: *Speaker Identification Error Rate for 8 APs, 26 MFCCs and 39 MFCCs*

Pop. Size	Gender (# of test utt)	8 APs	26 MFCCs	39 MFCCs
50	Female (1379)	28.06	31.91	30.75
	Male (1308)	25.00	22.40	22.55
	Average	26.53	27.16	26.65
100	Female (1104)	31.71	36.59	35.15
	Male (1093)	27.17	25.62	26.35
	Average	29.44	31.11	30.75
125	Female (1379)	33.65	38.95	37.49
	Male (1308)	30.04	27.37	27.98
	Average	31.85	33.16	32.74
250	Female (1379)	36.69	43.29	42.57
	Male (1308)	34.40	31.65	32.11
	Average	35.55	37.47	37.34

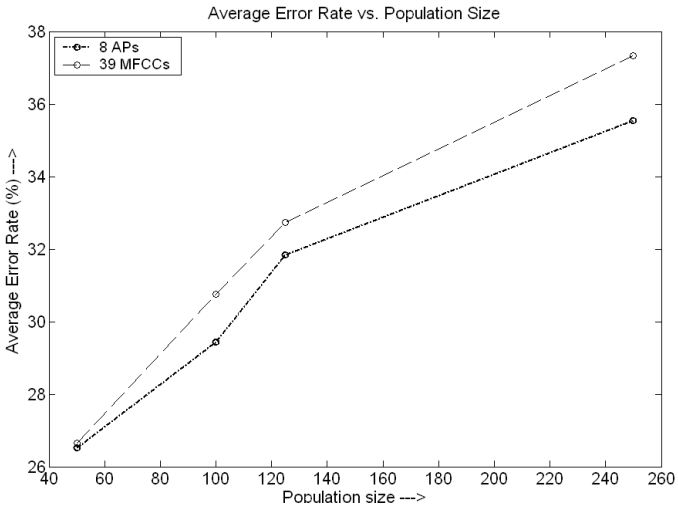

Figure 4: *Average Error Rate vs. Population Size*

We observe that our features give a significant improvement over the performance of the 26 MFCCs and 39 MFCCs for the female speakers. However for the male speakers, the MFCCs yield better performance. This trend is consistent for different population sizes. A plausible explanation is that females differ more in the degree of breathiness, and this information should be captured and well modeled by our source parameters. To avoid biasing the results, an unweighted average was taken since the number of female speakers is higher. On the average, the 8 APs have a comparable performance to the 26 and 39 MFCCs.

The average error rate increases with increasing population size (Figure 4). This is expected since the number of confusions increase for a larger population. The performance gap between the 8 APs and 39 MFCCs increases as the speaker population grows larger. In this paper, some of the identification error rates obtained are high, since we have performed the task on large populations of the same gender.

6. Conclusions and Future Work

In this paper, we have discussed a set of acoustic parameters that attempt to capture speaker-specific characteristics. Our feature set consists of only eight parameters and it has a comparable performance to the standard set of 26 or 39 MFCCs for text-independent speaker identification task. Work is in progress to add more parameters to our existing feature set in order to obtain more speaker-specific information. In particular, we are presently working on the automatic extraction of creakiness and nasalization. We plan to conduct more detailed experiments to test the accuracy and robustness of our feature set. We would also like to capture the temporal information of the parameters and would have to use a framework other than the GMMs. Additionally, we will explore a supervised approach to speaker identification since the relationship between acoustic properties and source/vocal tract information may change across classes of speech sounds.

7. Acknowledgements

This work was supported by NSF grant # BCS-0519256. We would like to thank Dr. Douglas Reynolds for the MIT-LL GMM System.

8. References

[1] J. Campbell, ``Speaker recognition: A tutorial," Proc. of the IEEE, vol. 85, pp. 1437--1462, Sept 1997.

[2] M. Plumpe, T.Quatieri, & D.Reynolds, “Modeling of the glottal flow derivative waveform with application to speaker identification”, IEEE Trans. Speech and Audio Proc., vol. 1, no. 5, pp. 569-586, Sept. 1999.

[3] D. Klatt and L. Klatt, “Analysis, synthesis and perception of voice quality variations among female and male talkers,” J. Acoust. Soc. Am. 87, 820-857, 1990.

[4] O. Deshmukh, C.Espy-Wilson, A. Salomon & J. Singh, “Use of Temporal Information: Detection of the Periodicity, Aperiodicity and Pitch in Speech," IEEE Trans. on Speech and Audio Proc.,” vol.13, pp. 776 – 786, September 2005.

[5] C. Espy-Wilson, “Articulatory strategies, speech acoustics and variability”, From Sound to Sense, June, 2004.

[6] C. Espy-Wilson, “Acoustic measures for linguistic features distinguishing the semivowels /wjrl/ in American English, J. Acoust. Soc. Am. 92, 1992, 736-757.

[7] Delattre, P. & Freeman, D, "A Dialect Study of American R's by X-ray Motion Picture," *Language*, 44, 29-68, 1968.

[8] G. Fant and M. Bavegard, “Parametric model of VT area functions: vowels and consonants,” (Stockholm), 38, 1-21, 1997.

[9] H. Takemoto, K. Honda, S. Masaki, Y. Shimada and I. Fijimoto, “Modeling of the inferior part of the vocal tract based on analysis of 3D cine-MRI data,” Proc. Autumn Meet. Acoust. Soc. Jpn., 2003, 281-282.

[10] D. Reynolds and R. Rose, “Robust text-independent speaker identification using Gaussian mixture speaker models,” IEEE Trans. Speech Audio Proc., vol. 3, no. 1, pp. 72–83, 1995.