ON THE INFLUENCE OF RATE, PITCH, AND SPECTRUM ON AUTOMATIC SPEAKER RECOGNITION PERFORMANCE

T.F. Quatieri, R.B. Dunn, and D.A. Reynolds
MIT Lincoln Laboratory, Lexington, MA, USA
{tfq,rbd,dar}@sstll.mit.edu

ABSTRACT
In this paper we study the influence of speech articulation rate, pitch, and spectrum on a GMM-based automatic speaker recognition system [2]. Using the high-quality sinusoidal transformation system [1], these factors are varied in a controlled manner and the effect on recognition performance evaluated. In general, there was found a larger loss in performance using modified speech for female than for male speakers due to greater feature dependence on spectral fine structure with increasing pitch. An important observation in this study is that certain transformations can dramatically alter the audibility of test data with little change in automatic recognition performance, particularly for male speakers. In addition, the influence of these rate, pitch, and spectral factors on recognition performance is important in order to understand the vulnerabilities of speaker recognition systems to speech modified for gain or false acceptance. For this purpose, we also investigate performance behavior when impostor (or target) speech alone is modified.

1 INTRODUCTION
Recently, there has been increasing interest in the performance of speaker recognition systems from modified speech [3, 4]. Little performance change was found by in an HMM-based recognizer, for example, when using synthetic speech from an all-pole noise-driven synthesis to produce the effect of whispered speech [4]. In this paper, motivated by this result, we the high-quality sinusoidal transformation system (STS) [1] to study the effect of a wide variety of speech modifications on a state-of-the-art GMM recognizer [2]. Transformations include change in articulation rate, pitch, and spectrum, as well as whispering.

In one class of recognition experiments, all test data, i.e., target and impostor trials, is modified to obtain a general understanding of sensitivity to modified speech. In altering articulation rate, little change in performance is obtained. On the other hand, pitch change leads to an increasing loss in performance with increasing pitch. Likewise, a similar sensitivity was found for the whispering transformation. This gender-dependence of the effect of pitch change or whispering is shown to correlate with the pitch-dependence of the mel-frequency feature used by the speaker recognizer; the sensitivity in the mel-frequency increases with increasing pitch with these transformations. Finally, in the first set of experiments, a spectral warping is invoked and shown to result in severe loss in recognition for both males and females. In a second class of recognition experiments, the target and impostor test data are modified separately, motivated by the goal of understanding speaker recognition vulnerability. Modification of the target data alone was found to work best in terms of performance by bringing the impostor speech closer to the background. Finally, informal listening tests are described, illustrating the degree of correlation between human- and machine-based speaker recognition for a variety of speech modifications.

2 SINUSOIDAL TRANSFORMATION SYSTEM
2.1 Analysis/Synthesis
The sinusoidal analysis/synthesis system is based on the premise that speech can be modeled as a sum of sinewaves with time-varying amplitudes, frequencies, and phases [1]. When the speech production model is introduced, both an excitation and a vocal tract contribution can be represented in the sinewave amplitudes and phases. Specifically, as in [1], the excitation function (quasi-periodic pulse train, noise, or combinations of pulses and noise) is represented by a sum of sinewaves with time-varying amplitudes $a_k(t)$ and frequencies $\omega_k(t)$, and let the vocal tract (combined with the glottal flow velocity waveform during voicing) be represented by a system function denoted by $H(\omega) = M(\omega)\exp[\imath\phi(\omega)]$. By combining the excitation and vocal tract amplitudes and phases, the representation can be written concisely as

$$s(t) = \sum_{k=1}^{K(t)} a_k(t) \exp[\imath \phi_k(t)]$$

where

$$A_k(t) = a_k(t) M[t, \omega_k(t)]$$

$$\theta_k(t) = \phi_k(t) + \Phi[t, \omega_k(t)]$$

$$= \int_0^t \omega_k(\tau) d\tau + \Phi[t, \omega_k(t)]$$

represent the amplitude and phase of the $k$th sinewave along the frequency trajectory $\omega_k(t)$.

2.2 Modification
The sinusoidal transformation system (STS) uses an analysis/synthesis based on the above sinewave model. Speech transformations stretch, compress, and scale sinewave frequency trajectories in time and frequency for time-scale and pitch modification, respectively. These transformations occur while preserving the original phase coherence (i.e., phase relations) or a speech-like coherence among sinewaves, thus preserving quality [1]. The algorithm can also perform these operations simultaneously. In addition, other transformations can also be simulated. The effect of whispering, i.e., an unvoiced (noise) excitation only, can be invoked by replacing the system phases $\Phi[t, \omega_k(t)]$ by uniformly distributed random variables on the interval $[-\pi, \pi]$. The spectral transformation of spectral warping can be obtained by stretching or compressing the system function amplitude and phase functions, $M[t, \omega]$ and $\Phi[t, \omega]$,
the frequency axis \( \omega \), thus simulating the effect of a longer or shorter vocal tract. In the following sections, we investigate the effect of these transformations on automatic speaker recognition performance.

3 BASELINE RECOGNITION

Speaker recognition experiments in this paper are performed using a state-of-the-art Gaussian mixture model universal background model (GMM-UBM) speaker verification system [2]. The primary components of the GMM-UBM system are front-end processing to extract 19 mel-cepstral features from mel-filter energies computed at a 10 ms frame interval, linear channel compensation in the form of cepstral mean removal and KASTA, and target and background model estimation. In addition, 19 delta cepstra coefficients (i.e., the slope of a straight-line fit to feature vectors over time) are computed over a 5 frame interval and appended to produce a 38 dimensional feature vector. Likelihood ratio scores are generated from test utterances of target and impostor (also referred to as “non-target”) speakers. The evaluation database is from the NIST eval2000 Switchboard-2 (Phase 1 and 2) database recorded with both carbon button and electret handsets [3]. The background model is trained from Switchboard-2 (Phase 3). No score normalization is used in this study. For the female gender, 504 target speakers are used with 3156 target test trials (cases where the speaker in the test file and model are the same) and 3027 impostor test trials (cases where the speaker in the test file and model are different); for the male gender, 422 target speakers are used with 2064 target test trials and 3034 impostor test trials. The performance of this baseline system is computed by separately pooling all target trial scores and all impostor trial scores and sweeping out a speaker-independent threshold over each set to compute the system’s miss and false alarm probabilities. The tradeoff in these errors is then plotted as a detection error tradeoff (DET) curve.

Experimental focus in this paper is on the “mismatched condition” in the sense that the background and training data are mismatch and the test data (target and/or impostor) are modified with STS. To obtain a performance reference for these experiments, however, we first obtain the recognition performance with sinewave analysis/synthesis without modification. The frame interval used in analysis/synthesis, as well as in modification, is 10.25 ms to avoid synchrony with the 10.0 ms GMM frame interval and thus to avoid a bias toward the original speech spectrum. Figure 1 shows that analysis/synthesis only does not change DET performance of the recognizer for males (thick solid versus thick dashed line). A similar characteristic is also seen for females as well, but with a slight decrease in performance. In the following section, we look at performance changes relative to this reference for a variety of transformations on all test data, i.e., both target and impostor test trials. In Section 5, we pursue out performance for target and impostor test trials.

4 MODIFICATION OF ALL TRIALS

4.1 Articulation Rate Change

Our first transformation is time-scale modification, i.e., a change in the articulation rate. For this transformation, we expect little change in performance because the stretching or compressing of the sinewave amplitude and phase functions does not effect the local spectral amplitude. Indeed, Figure 1 (thin solid line) shows that articulation rate change gives a small loss in performance for both male and female speakers for the case of timescale expansion by a factor of two. A similar loss occurs for an equivalent compression. The loss, although small, indicates that articulation rate is reflected in recognition performance. This may be due to time-scale modification imparting greater or lesser stationarity to the synthetic speech than in the original so that both the mel-cepstrum and the delta-cepstrum are altered, particularly during event transitions and foreground movement. Although there is some gender difference in recognition performance, this difference is minor, as expected, because stationarity is seemingly not gender-dependent. In informal listening, the loss in recognition performance, however, does not correspond to a loss in aural speaker identifiability. The speaker characteristic appears
to be preserved; only the speaking rate is modified.

4.2 Pitch Change

Common knowledge is that a change in pitch should not affect recognition performance because pitch does not measurably alter the mel-filter energies. To test this hypothesis, we first performed a pitch change by modifying the pitch contour to take on a fixed, monotone pitch. For the case of 60 Hz monotone pitch, Figure 1 shows that the common knowledge of the unimportance of pitch appears to hold for male speakers; but for female speakers, a significant loss in performance occurs. A similar result is found when the time-varying pitch is replaced by a monotone 100 Hz pitch, as well as when pitch contours of the test data are shifted and scaled by numerous values to impart a change in the pitch mean and variance. In all cases of pitch modification, the male recognition performance changes negligibly, while female performance changes significantly. In informal listening, aural speaker identifiability is essentially lost with these modifications; for male speakers, this loss by human listeners is not consistent with negligible performance loss by automatic recognition.

To understand the gender-dependent performance loss with pitch modification, we need to understand how the mel-spectrum features change with a change in pitch. Toward this end, we compare mel-filter energies (from which the mel-spectrum is derived) from a high-pitched female and a low-pitched male. Figure 2 shows the speech log-spectrum and mel-filter energies for one frame from each speaker. The mel-filters are those used in the GMM recognition front-end. It is seen that there is clear harmonic structure, most notably in the low frequencies of the mel-filter energies of the female speech. In practice, we see a continuum in the degree of harmonic structure in going from high-pitched to low-pitched speakers. Mel-filter energies do, therefore, contain pitch information. Furthermore, we speculate that the delta mel-spectrum contains, in addition, pitch motion. For high-pitched speakers, a change in pitch will modify the mel-spectrum and thus create a "mismatch condition" between the training and test data.

The result of combining the transformations of rate and pitch change (time-scale expansion by two and a 60 Hz monotone pitch) is also shown in Figure 1 (thick dashed-dotted line). Here we see that performance is essentially equivalent to that of the monotone pitch transformation, the change in mel-spectral coefficients being dominated by the pitch change.

Figure 2: Comparison of log-spectrum and log-mel energies for female (a) and male (b) speakers.

Figure 3: Comparison of narrowband spectrograms of whispered synthetic speech (upper panel) and its original counterpart (lower panel).

4.3 Whispering

The next transformation is that of converting the excitation to white noise to produce the effect of whispered speech. It was stated earlier that this effect can be created with STS by replacing the sine-wave system phases by uniformly distributed random phases on the interval $[-\pi, \pi]$. Figure 3 illustrates that the formant information remains essentially intact with this sine-wave phase randomization. The recognition result for this transformation on all test trials of our corpus is given in Figure 1 (thin dashed line). We see that the performance change is very similar to that of the monotone pitch transformation: a large loss for female speakers and a negligible loss for male speakers; indeed, for each gender, the DET curves are close for the two very different transformations. We speculate that, as with pitch transformations, this gender-dependent performance loss is due to the increased importance of fine structure in the mel-spectrum with increasing pitch; as seen in Figure 3, pitch is completely removed with phase randomization, while gross spectral structure is preserved. For this transformation, there is a large loss in aural speaker identifiability, as there occurs with naturally whispered speech; yet there is little performance loss for male speakers in automatic speaker recognition.

4.4 Spectral warping

The final transformation invokes a warping of the spectral magnitude $M(t, \omega)$ and phase $\Phi(t, \omega)$ (defined in Section 2.1) which corresponds to a change in length of the vocal tract. With this transformation, we expect a large loss in recognition performance because we are displacing the speech formants and modifying their bandwidths, thus creating a severe "mismatch condition" between the training and test data. Figure 1 (thick dotted line) confirms this expectation, showing a very large loss in performance for both males and females; in fact, the performance is nearly that of a system with random likelihood ratio scores. For this transformation, there also occurs a significant loss in aural speaker identifiability. A similar result was obtained with a 20% spectral compression. Figure 1 also clearly shows the relative importance of spectrum versus prosody (articulation rate and pitch) and whispering for automatic recognition.

5 TARGET VERSUS IMPOSTER MODIFICATION

In this section, we investigate recognition performance when the transformations of the previous section are applied separately to parsed target and impostor trials. The result of transforming the target trials only is shown in the left panel of Figure 4 for male speakers. We see that generally there is a greater degradation in performance (i.e., for a fixed
miss rate, there is a greater false alarm rate than for the case where all trials are transformed (as in Figure 1). This is likely because the target mel-cepstrum is moved away from the target models while the imposter mel-cepstrum is unchanged. On the other hand, there is improved performance when the imposter is transformed, likely because the mel-cepstrum of the imposter is moved closer to that of the background model, thus reducing the false alarm rate for a fixed miss rate. Therefore, the recognition system is less vulnerable to an imposter speaker who attempts to disguise his/her voice through the above transformations, or an “impostor” with spectra of the target speaker but not the target’s excitation or prosody.

6 DISCUSSION AND FUTURE WORK

An important aspect of our study is the insight gained into the degree of correlation of the loss in speaker identifiability by human and by machine when speech is modified. For example, we have seen that for males, pitch change and whispering can dramatically alter aural speaker identifiability, while essentially maintain automatic recognition performance when all test trials are modified. For female speakers, on the other hand, these transformations lead to recognition loss by both human and machine; for the machine, the original pitch, removed in these transformations, may be a key recognition component. Conversely, there are cases where automatic recognition can degrade but without change in aural speaker identifiability, as with articulation rate changes. Nevertheless, it is clear that the human is using something beyond spectro-temporal primary features currently used in state-of-the-art automatic speaker recognition systems. This observation motivates the search for non-spectral features in automatic recognition.

The transformations in this paper are representative of speech modifications but not exhaustive. For example, removal of spectral tilt may be of interest. Spectral tilt corresponds to the shape of the glottal flow waveform during voicing and is associated with different voice types. Removal of spectral tilt, achieved by removing the spectral trend in the system function magnitude $M(f)$ in the sine-wave model of Section 2, gives significant performance loss in automatic recognition for both male and female speakers. In addition, noticeable loss in aural speaker identifiability is found.

Finally, we noted that a motivation for this study is to understand vulnerabilities of speaker recognition systems. For the transformations in this paper, we have seen, from parsing results for target and imposter trials, that automatic recognition becomes less vulnerable to attacks by imposters who use transformations as a vehicle for disguise or mimicry. Nevertheless, further work is needed to better understand vulnerabilities of automatic speaker recognition to modified speech.

Acknowledgement: The authors thank Fred Goodman of Autometric for many useful discussions during this effort.

REFERENCES


Switchboard-2 Phase 1, 2, and 3.