Speech Style and Speaker Recognition: a Case Study

Marco Grimaldi1,2, Fred Cummins1

1School of Computer Science and Informatics, UCD, Dublin
Marco.Grimaldi@ucd.ie, Fred.Cummins@ucd.ie
2FBK, via Sommarive 18, I-38100 Povo (Trento)
grimaldi@fbk.eu

Abstract

This work presents an experimental evaluation of the effect of different speech styles on the task of speaker recognition. We make use of willfully altered voice extracted from the CHAINS corpus and methodically assess the effect of its use in both testing and training a reference speaker identification system and a reference speaker verification system. In this work we contrast normal readings of text with two varieties of imitative styles and with the familiar, non-imitative, variant of fast speech. Furthermore, we test the applicability of a novel speech parameterization that has been suggested as a promising technique in the task of speaker identification: the pykfec. The experimental evaluation indicates that both the reference verification and identification systems are affected by variations in style of the speech material used, especially in the case that speech is also mismatched in channel. Our case studies also indicate that the adoption of pykfec as speech encoding methodology has an overall favorable effect on the systems accuracy scores.

Index Terms: speaker recognition, speech style, instantaneous frequencies

1. Introduction

Speaker identification and verification techniques aim to provide procedures that allow the robust recognition of speakers in a variety of speaking conditions and recording scenarios. Researchers (e.g.: [9, 10, 16, 18]) working in the field often divide the principal factors that affect the robustness of automatic speaker recognition systems into two broad categories: variation in communication channel and variation in the speaker’s voice.

The experimental control and manipulation of the communication channel is relatively straightforward, even though the remedies for distortion may not be obvious. There has thus been a great deal of empirical work into techniques for alleviating various problems associated with channel variation. Researchers working in this area have demonstrated that several different approaches to channel compensation can be applied to mitigate the loss in recognition accuracy, e.g.: cepstral mean subtraction [14], RASTA processing [7], feature warping [11] and fusion techniques [17].

Variability in a person’s voice may include the accidental or intentional modification of speech. This may be due to physiological conditions (cold, stress, etc) or intentional disguise (e.g.: by the willful adoption of distinct speaking styles such as whispering or fast rate speech, etc) [16]. In considering the unintentional modification of speech, physiological conditions such as seasonal effects are usually normalized [2] by recording speakers in well-separated sessions, often months apart. On the other hand, little is known about the problems which might be raised for speaker recognition as result of intentional disguise or stylistic variation. These hurdles present themselves with particular force in the forensic situation, where there is little or no control over the recording conditions, speakers may be under extreme stress, and disguise may be used [1, 16]. In the area of speaker verification, some recent research work (e.g.: [9, 10]) shows that automatic speaker verification systems are badly affected by the use of disguised voice or by re-synthesis of the client speech.

In this work, we stress the use of willfully altered voice and methodically assess the effect of its use in both testing and training of a reference speaker identification system. We here refer to willfully altered voice as a volitional alteration deviation from normal voice by adoption of an operationally defined speaking style. The speech material is extracted from the CHAINS corpus. In the present work we contrast normal readings of text with three varieties of willfully altered voice, repetitive synchronous imitation, synchronous speech and fast rate speech.

This work also tests the applicability of an alternative speech parameterization that has been suggested to be robust with regard to speech style variations [6]. The pykfec - pykogram frequency estimate coefficients - are derived from an AM/FM approximation of the input signal, do not require any channel compensation schema and compare favorably with standard MFCC and RASTA features in speaker identification [6]. The results obtained parameterizing speech using pykfec are systematically compared with results obtained using standard MFCCs as speech parameters. Section 3 provides details of both pykfec and MFCC parameterization of speech as adopted in this work.

The experimental evaluation consists of two main phases: First, a generic identification system (implementing a hard-match among the speakers) is tested varying speech material (style and channel) and speech encoding (MFCCs and pykfec). Then, the effect of channel and style variations using both pykfec and MFCCs is assessed in the open-set speaker verification scenario.

2. Willfully altered speech

The speech material used in this work is extracted from the CHAINS corpus. The corpus contains the recordings of 36 speakers obtained in two different sessions with a time separation of two to three months. Across the two sessions, each speaker provided recordings in six different speaking styles. Each speaking style has a clear operational definition: speakers were recorded in a distinct, experimentally controlled, speaking...
condition and the produced speech is labelled accordingly. Full details of all the six speaking styles and further details about the corpus such as speaker gender distribution, dialectal origin and the recorded text material can be found in [3].

In this work, four speaking styles are used:

— Normal speech (NORM): subjects read prepared text drawn from a set of short fables and sentences. No constraint on rate or manner was imposed. The speech material belongs to the first recording session of the corpus.

— Synchronous Speech (SYNC): two subjects read a prepared text (fable/sentence) in synchrony. The resulting speech is typically at a relatively slow rate, and does not sound markedly different from normal speech [4]. The corresponding speech material also belongs to the first recording session of the corpus.

— Repetitive Synchronous Imitation (RSI): subjects repeat a short-phrase they hear in a continuous loop. They hear the sum of their own production and the model phrase. The effect of this manipulation is to produce a very strong sense of mismatch between one’s own speech and that of the target, and the result is an automatic adjustment of one’s own speech to make it as similar as possible to the target [3]. The material belongs to the second recording session of the corpus.

— Fast Speech (FAST): finally, fast speech was used, as this is a well-known speech style, easily elicited, and not imitative in nature. The FAST recordings were also from the second recording session.

3. Speech Parameterization

In this work two different speech parameterizations are considered: standard MFCCs and pykfec - pyknogram frequency estimates frequency coefficients.

3.1. AM-FM: pykfec

In a recent work [6], pykfec have been proposed as an alternative parameterization of the speech signal for speaker recognition purposes. This new parameterization is based on the AM-FM (e.g.: [5, 8, 12]) approximation of the input signal, does not require any channel compensation and compares favorably with MFCCs and RASTA in speaker identification [6]. As reported in [6], pykfec can be extracted from the input signal by passing the speech through a filterbank of Gabor filters equally spaced along the frequency axis and then performing demodulation of the band-passed signal (adopting a multi-band demodulation schema as discussed also in [12]).

In this work we characterize the input signal by using 40 linearly spaced Gabor filters between 0 Hz–4000 Hz, with constant bandwidth of 106 Mel (setup-3 in [6]), using a short-time window of about 25 msec and an overlapping-window of about 12.5 msec.

3.2. MFCC

MFCCs are extracted using 40 triangular filter spaced between 0 Hz–4000 Hz. The window length is again set to 25 msec with an overlap between successive windows of one half its value, to ensure homogeneous sampling between the two approaches (MFCC and pykfec) for the same input files. The zeroth cepstral coefficient is not used in the Mel-frequency cepstral fea-

1referred to as SOLO in the CHAINS corpus.
2A reference implementation of pykfec extraction is available at http://chains.ucd.ie/ftpaccess.php.

4. Speaker Recognition: Methodology

The speaker’s model is obtained using a generic Gaussian Mixture Model induction algorithm: the mixture centers and variances are first estimated using a k-mean clustering mixture algorithm; in a subsequent phase, the Gaussian models and their associated weights are refined using the expectation maximization algorithm (EM) [15].

The effect of training and testing the GMM speaker recognition system while varying the speech style is assessed in both a closed-set speaker identification task and in an open-set speaker verification scenario.

First, the generic identification system (implementing a hard-match among the speakers) is tested using test and train material that is matched in both style and channel. This set of measures provide the baseline score: the upper limit that the system can reach across the styles. Then the same classifier is tested with training and test materials mismatched in style, but matched within one recording session. Finally, the identification system is tested with training and test materials mismatched for both style and recording session. This set of experiments constitutes the hardest task and best approximates a real case scenario. 64 GMMs per speaker are used in the modeling phase of the identification task throughout.

The effect of channel and style variations using both pykfec and MFCCs is also assessed in the open-set speaker verification scenario. The verification system uses a Universal Background Model (UBM – 1024 GMMs) to approximate the speaker population, while the claimant model is estimated using 64 GMMs. The performance of the system is evaluated providing separate train, test, UBM and impostor sets. The testing procedure involves the selection of a claimant speaker and all the available impostors. After all the selected speakers are tested, a new claimant is extracted from the pool of available speakers and the same procedure repeated. As in the previous case, the verification system is trained and tested first with speech matched in both style and channel, then for speech matched in channel but elicited using different styles and finally for speech mismatched in both channel and style.

The accuracy of the identification/verification system is evaluated with 10 second test utterances and is expressed as the mean of ten runs, each run having a different random initialization of the GMMs. The error of the recognition score is calculated as twice the standard deviation of the mean, corresponding to a confidence interval of about 95%. In the verification task, the accuracy of the system is defined as EER (Equal Error Rate).

4.1. Datasets

Three different training sets are used in the evaluation. Two training sets are obtained from the first recording session of the corpus (NORM, SYNC), about 60 seconds of speech per speaker. The remaining one uses FAST speech from the second recording session as modified speech material, about 60 seconds of speech per speaker.

To test the trained models four test sets are used: two extracted from the first recording session (NORM, SYNC), using speech not previously selected for training, about 25 seconds of
speech per speaker. The other two test sets are selected using material from the second recording session of the corpus: FAST, RSI - about 25 seconds of speech per speaker.

In the speaker identification task, the train and test sets are made up by all the available speakers (36). In the verification case, speakers are divided into 3 groups: 8 speakers act as claimants, 24 speakers are used for UBM modeling, the remaining 4 speakers form the impostor set. Speakers are selected randomly.

5. Results

5.1. The Identification Task

Figure 1 shows the accuracy curves for the reference identification system varying speech encoding (pykfec, MFCCs) and varying speech style and number of features.

![Figure 1: Accuracy of a generic identification system trained and tested using speech material mismatched in both style and channel.](image)

The results in Figure 1(a) indicate that the performance is stable at about 100%, regardless of the number of parameters and encoding adopted, when the recorded material used in training and testing the GMMs is within the same channel and speech style. This is not surprising since the speech material used is extracted from the same recording session. On the other hand, Figure 1(b) shows that when different styles are used in training and testing the algorithm - while maintaining the selected material within channel and thus using speech material belonging to the same recording session - a decrease of accuracy in the performance of the same system is registered across all the different scenarios tested. In terms of absolute performance both MFCC and pykfec show similar accuracy scores.

Figure 2 shows the accuracy curves for the reference identification system varying speech encoding (pykfec, MFCCs), using speech material mismatched in both style and channel.

![Figure 2: Accuracy of a generic identification system trained and tested using speech material mismatched in both style and channel.](image)

5.2. The Verification Task

Table 1 summarizes results obtained across styles adding delta features (Δ) to the base descriptors. Results are grouped by channel: the top three lines show results obtained using speech styles belonging to the same recording session (hence channel invariant); the bottom part shows results obtained using speech material mismatched in both channel and style.

<table>
<thead>
<tr>
<th>Train</th>
<th>Test</th>
<th>MFCC 15 [%]</th>
<th>MFCC 15+Δ [%]</th>
<th>pykfec 20 [%]</th>
<th>pykfec 20+Δ [%]</th>
</tr>
</thead>
<tbody>
<tr>
<td>NORM</td>
<td>SYNC</td>
<td>100 ± 1</td>
<td>100 ± 1</td>
<td>98 ± 1</td>
<td>99 ± 4</td>
</tr>
<tr>
<td>SYNC</td>
<td>NORM</td>
<td>96 ± 1</td>
<td>97 ± 1</td>
<td>97 ± 1</td>
<td>97 ± 1</td>
</tr>
<tr>
<td>FAST</td>
<td>RSI</td>
<td>92 ± 1</td>
<td>92 ± 1</td>
<td>91 ± 1</td>
<td>93 ± 1</td>
</tr>
<tr>
<td>NORM</td>
<td>FAST</td>
<td>85 ± 2</td>
<td>88 ± 2</td>
<td>84 ± 1</td>
<td>86 ± 2</td>
</tr>
<tr>
<td>FAST</td>
<td>NORM</td>
<td>85 ± 2</td>
<td>74 ± 2</td>
<td>71 ± 3</td>
<td>85 ± 2</td>
</tr>
<tr>
<td>FAST</td>
<td>SYNC</td>
<td>60 ± 3</td>
<td>74 ± 2</td>
<td>73 ± 1</td>
<td>78 ± 2</td>
</tr>
<tr>
<td>SYNC</td>
<td>FAST</td>
<td>69 ± 2</td>
<td>79 ± 1</td>
<td>71 ± 3</td>
<td>80 ± 2</td>
</tr>
<tr>
<td>NORM</td>
<td>RSI</td>
<td>73 ± 2</td>
<td>77 ± 3</td>
<td>74 ± 2</td>
<td>84 ± 2</td>
</tr>
<tr>
<td>SYNC</td>
<td>RSI</td>
<td>56 ± 3</td>
<td>65 ± 2</td>
<td>75 ± 3</td>
<td>84 ± 1</td>
</tr>
</tbody>
</table>

Table 1 indicates that the addition of Δ features is beneficial to system performance for both MFCCs and pykfec, especially when the speech material is mismatched both in style and channel. In this case, it indicates a benefit in adopting pykfec over MFCCs, in terms of gain in prediction score.
to handle differences in speaking style well when no variations in channel are present. When the speech material is also mismatched in channel, the system accuracy varies across the different scenarios considered in much the same way as before. As in the previous case, Table 2 indicates that the adoption of pykfec as a speech encoding methodology has a positive effect on system accuracy (defined as 1-EER).

Figure 3 shows sample DET curves obtained using different channels, styles and speech encodings in training and testing the verification system.

**6. Conclusions**

In this work we have analyzed the effect of willfully altered speech, as provided by the CHAINS corpus, in the context of speaker recognition. Methodically, we assessed the effect of different speech styles in testing and training a identification system and a verification system also varying the speech encoding adopted (MFCCs, pykfec, number of features). The evaluation indicates that both the reference verification and identification systems are similarly affected: when speech is matched in channel are present. When the speech material is also mismatched only in style, a relatively small loss in accuracy is recorded. When the speech material is mismatched in both channel and style, the accuracy of the two systems varies greatly across the different scenarios employed and the accuracy results indicate that the adoption of pykfec as a speech encoding methodology has a positive overall effect on the accuracy scores.

**7. Acknowledgments**

This work is supported by Science Foundation Ireland grant No. O4/IN3/I568 to the second author.

**8. References**


