Speech Style and Speaker Recognition: a Case Study

Marco Grimaldi, Fred Cummins

School of Computer Science and Informatics, UCD, Dubin

Marco.Grimaldi@ucd.ie, Fred.Cummins@ucd.ie

Abstract

This work presents an experimental evaluation of the effect of different speech styles on the task of speaker identification. Although the informal notion of a speaking style does not readily translate into operational procedures for eliciting speech in one style or another, 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 *pyknogram frequency estimate coefficients - pykfec*. The experimental evaluation indicates that both the reference verification and identification systems are affected by variations in style of the speech material used. Our case studies also indicates 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, 11, 17, 19]) 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 variations 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 [15], RASTA processing [7], feature warping [12] and fusion techniques [18], feature mapping [14].

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) [17]. 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 recording conditions, speakers may be under extreme stress, and disguise may be used [1, 17]. In the area of speaker verification, some recent research work (e.g.: [9, 11]) 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 or deviation from normal voice by adoption of an *operationally* defined speaking style. The speech material is extracted from the CHAINS corpus [3]. 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 - pyknogram 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[1]): subjects read prepared text drawn from a set of short fables and sentences. No constraint on rate or manner was imposed. Recordings were of very high quality (professional studio). Speech material recorded in this way is referred to as NORM, and 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): this procedure was originally developed for second language pedagogy [10]. 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 ([5, 8, 13]) 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[2] (adopting a multi-band demodulation schema as discussed also in [13]).

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 number of coefficients used for identification is not selected *a priori*: it is evaluated experimentally (Section 5) in order to maximize the speaker recognition rate. 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 feature vector, while the values of the coefficients are normalized using cepstral mean removal (e.g.: [15, 16]), in order to compensate for the different recording channels used to train and subsequently test the induction algorithm.

[1] referred to as SOLO in the CHAINS corpus.

[2] A 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 algorithm; in a subsequent phase, the Gaussian models and their associated weights are refined using the expectation maximization algorithm (EM)[16].

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) 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 for both style and channel, then for speech matched for 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 1-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 (SOLO, SYNC), about 60 seconds of speech per speaker, from 36 speakers. The remaining one uses FAST speech from the second recording session as modified speech material, about 60 seconds of speech per speaker, 36 speakers.

To test the trained models four test sets are used: two ex-

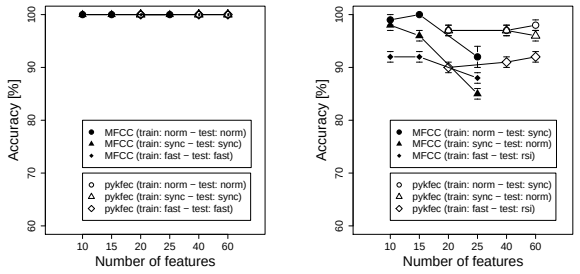
tracted from the first recording session (SOLO, SYNC), using speech not previously selected for training, about 25 seconds of speech per speaker. The other two (2) test sets are selected using material from the second recording session of the corpus: FAST, RSI - about 25 seconds of speech per speaker.

All the sets are used for speaker identification. 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.



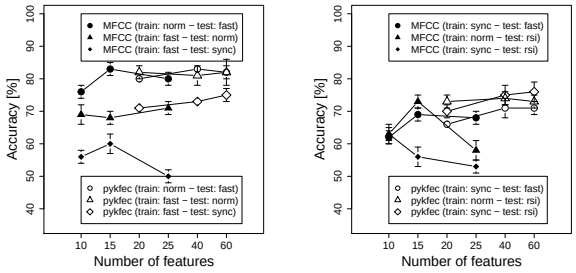
(a) (b)

Figure 1: Accuracy of a generic identification system trained and tested using: (a) speech material matched in style and channel; (b) speech material matched in channel and mismatched in style.

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 in training and testing the system is extracted from the same recording session. On the other hand, Figure 1(b) shows that when different styles are used in testing and training the algorithm - while maintaining the selected material within channel - 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(a),(b) clearly indicates that the system accuracy is badly affected when speech mismatched both in style and channel is used in training and testing the algorithm, with actual accuracy scores varying greatly depending on the style adopted in the training and testing phase and the number of features and the speech encoding adopted, e.g.: training the algorithm using NORM speech and testing using RSI, the system scores $(58 \pm 3)\%$ – 25 MFCCs; training using NORM speech and testing using FAST, the algorithm scores $(80 \pm 2)\%$ – 25 MFCCs. Adopting 20 *pykfec* to encode the signal, the system scores



(a) (b)

Figure 2: Accuracy of a generic identification system trained and tested using speech material mismatched in both style and channel.

$(74 \pm 2)\%$ training the algorithm using NORM speech and testing using RSI.

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 1: Accuracy scores obtained using the same generic identification system varying encoding, trainin) and testing speech material.

Train	Test	MFCC 15 [%]	MFCC 15+Δ[%]	pykfec 20 [%]	pykfec 20+Δ [%]
NORM	SYNC	100 ± 1	100 ± 1	98 ± 1	99 ± 1
SYNC	NORM	96 ± 1	97 ± 1	97 ± 1	97 ± 1
FAST	RSI	92 ± 1	92 ± 1	91 ± 1	93 ± 1
NORM	FAST	83 ± 2	88 ± 2	83 ± 1	89 ± 2
FAST	NORM	68 ± 2	74 ± 2	81 ± 3	85 ± 2
FAST	SYNC	60 ± 3	74 ± 2	73 ± 1	78 ± 2
SYNC	FAST	69 ± 2	79 ± 1	71 ± 3	80 ± 2
NORM	RSI	73 ± 2	77 ± 3	74 ± 2	84 ± 2
SYNC	RSI	56 ± 3	65 ± 2	75 ± 3	84 ± 1

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.

5.2. The Verification Task

Table 2 shows the results obtained training and testing the generic verification system varying speaking style. In this section we use the term 'training material' to indicate speech material used to train the claimant and the UBM models; similarly, 'test material' indicates the speech material used to test claimant and the impostors. Results are grouped by channel: in the top part we report results obtained considering speech matched in style and channel; the middle part shows results obtained with speech mismatched in style but not channel, while the bottom part shows results mismatched in both channel and style.

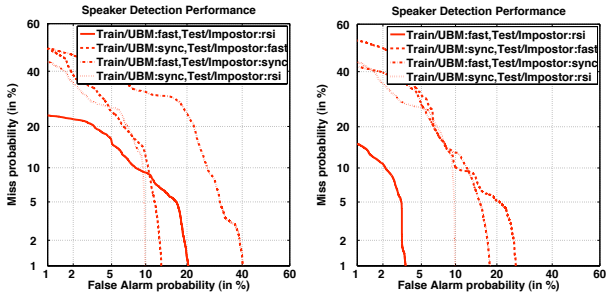
As in the case of closed-set identification, the system seems

Table 2: Accuracy scores obtained using the same generic verification system varying encoding, training and testing claimant speech material, UBM and impostor (Imp.) speech material.

Train	Test	UBM	Imp.	MFCC 15+Δ [%]	pykfec 20+Δ [%]
NORM	NORM	NORM	NORM	98 ± 2	100 ± 1
SYNC	SYNC	SYNC	SYNC	99 ± 1	100 ± 1
SYNC	NORM	SYNC	NORM	96 ± 1	99 ± 1
FAST	RSI	FAST	RSI	91 ± 1	96 ± 1
SYNC	FAST	SYNC	FAST	89 ± 1	90 ± 1
SYNC	RSI	SYNC	RSI	90 ± 1	92 ± 1
FAST	SYNC	FAST	SYNC	81 ± 4	88 ± 2

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 encoding in training and testing the verification system.



(a) MFCC (b) pykfec

Figure 3: Sample DET curves of the generic verification system trained and tested using different speech material.

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. The evaluation indicates that both the reference verification and identification systems are affected by variations in the style of speech used in testing and training. When the speech material is mismatched in both channel and style, the accuracy of the two systems varies predictably 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

[1] J.F. Bonastre, F. Bimbot, L.J. Boe, J. P. Campbell, D. A. Reynolds, and I. Magrin-Chagnolleau. Person authentication by voice: A need for caution. In *Proceedings of Eurospeech 2003*, Genova, September 2003.

[2] R. Cole, M. Noel, and V. Noel. The cslu speaker recognition corpus. In *Proceedings of the 5th International Conference on Spoken Language Processsing (ICSLP'98)*, Sydney, Australia, 30th November-4th December 1998.

[3] F. Cummins, M. Grimaldi, T. Leonard, and J. Simko. The chains corpus: Characterizing individual speakers. In *Proceedings of SPECOM'06*, pages 431–435, St. Petersburg, RU, 2006.

[4] Fred Cummins. Practice and performance in speech produced synchronously. *Journal of Phonetics*, 31(2):139–148, 2003.

[5] D. Dimitriadis and P. Maragos. Robust energy demodulation based on continuous models with application to speech recognition. In *Proceedings of Eurospeech 2003*, pages 2853–2856, 2003.

[6] M. Grimaldi and F. Cummins. Speaker Identification Using Instantaneous Frequencies. *Audio, Speech, and Language Processing, IEEE Transactions on*, 16(6):1097–1111, 2008.

[7] H. Hermansky, N. Morgan, and A. Bayya.and P. Kohn. Rasta-plp speech analysis technique. *IEEE International Conference on Acoustics, Speech, and Signal Processing, 1992. ICASSP-92.*, 1:121–124, 23-26 Mar 1992.

[8] C.R. Jr Jankowski, T.F. Quatieri, and D.A. Reynolds. Measuring fine structure in speech: Application to speaker identification. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '95)*, pages 325–328, 1995.

[9] S.S. Kajarekar, H. Bratt, E. Shriberg, and R. de Leon. A study of intentional voice modifications for evading automatic speaker recognition. In *Proceedings of the IEEE Odyssey 2006 Speaker and Language Recognition Workshop*, San Juan, Puerto Rico, 2006.

[10] H. Larson. Experiences of large scale implementation of speech analyzing tools in learning Swedish as second language. In *MATISSE-ESCA/SOCRATES Workshop on Method and Tool Innovations for Speech Science Education*. ISCA, 1999.

[11] J. Lindberg and M. Blomberg. Vulnerability in speaker verification - a study of technical impostor techniques. In *Proceedings of EUROSPEECH'99*, pages 1211–1214, Budapest, Hungary, September 1999.

[12] J. Pelecanos and S. Sridharan. Feature warping for robust speaker verification. In *In Proceedings of 2001: A Speaker Odyssey, The Speaker Recognition Workshop*, pages 213–218, Crete, Greece, 2001.

[13] A. Potamianos and P. Maragos. Speech formant frequency and bandwidth tracking using multiband energy demodulation. *Journal of the Acoustic Society of America*, 99:3795–3806, 1996.

[14] D.A. Reynods. Channel robust speaker verification via feature mapping. In *IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings. (ICASSP03)*, volume 2, pages 53–56, April 2003.

[15] D.A. Reynolds. Experimental evaluation of features for robust speaker identification. *IEEE Transactions on Speech and Audio Processing*, 2(3):639–643, 1994.

[16] D.A. Reynolds and R.C. Rose. Robust text-independent speaker identification using gaussian mixture speaker models. *IEEE Transactions on Speech and Audio Processing*, 3(1):72–83, 1995.

[17] Robert D. Rodman. Computer recognition of speakers who disguise their voice. In *Proceedings of ICSPAT 2000*. 2000.

[18] C. Senac and E. Ambikairajah. Audio indexing using feature warping and fusion techniques. In *IEEE 6th Workshop on Multimedia Signal Processing*, pages 359–362, September 2004.

[19] S.K. Taseer. Speaker identification for speakers with deliberately disguised voices using glottal pulse information. In *9th International Multitopic Conference, IEEE INMIC 2005*, pages 1–5, 2005.