EFFECTIVENESS OF SHORT-TERM PROSODIC FEATURES FOR SPEAKER VERIFICATION

Iker Luengo, Eva Navas, Inmaculada Hernáez, Jon Sanchez, Ibon Saratxaga, Iñaki Sainz

Aholab – Escuela Técnica Superior de Ingeniería.
Universidad del País Vasco – Euskal Herriko Unibertsitatea
Urkijo zum. z/g 48013 Bilbo
ikerl@bips.bi.ehu.es, eva.navas@ehu.es, inma.hernaez@ehu.es, jon.sanchez@ehu.es,
ibon.saratxaga@ehu.es, inaki@bips.bi.ehu.es

ABSTRACT

In this work a traditional MFCC based speaker verification system is combined with a prosody based one to determine whether simple short-term prosodic information is useful for improving current state-of-the-art ASV. The traditional speaker verification system based in spectral information has an EER of 3.85% when using 1024 mixtures. The prosody based system uses short-term intonation and energy information and achieves an EER of 23.93% with 128 mixtures. After applying LDA and fusing those scores, a final EER of 3.84% is achieved. This result does not show a significant improvement when compared with the result of the traditional speaker verification system.

1. INTRODUCTION

The use of long-distance transactions has become very popular in the last few years: shopping through the Internet, web based bank transactions, restricted access to secure areas in remote computers... All these systems need some kind of authentication procedure, in order to verify the users' identity. Most of them use password-type authentication, but passwords may be forgotten or stolen.

Nowadays biometric authentication is the best alternative. Biometric characteristics can't be lost nor forgotten, and are very difficult to imitate. This kind of authentication can already be seen in different applications: Laptops with fingerprint controlled access and hand geometry based access to certain buildings are some examples.

The growing interest on biometrics based automatic identification systems is reflected in the increase of biometric verification system contests, as the Fingerprint Verification Competition [1] or the NIST Speaker Recognition Evaluations [2], where new algorithms and methods are being proposed in order to improve current results. Automatic Speaker Verification (ASV) is not an exception. Most of the current implementations use some kind of spectral envelope features to parameterize the voice (MFCC, LPCC...), achieving a great performance [3] [4] [5]. But recent researches are trying to include prosodic information into the system, in order to reduce error rates.

Speech prosody refers to the intonation, energy and rate of the speech. It is well known that these features are characteristics of each person, so that they carry information about the speaker. Furthermore, prosody is uncorrelated with the spectral envelope shape. Therefore, supposedly adding prosodic features to the already used spectral features may lead to an improvement in the system's performance.

Most of the works in this area focus on the use of some kind of long-term prosodic information [6] [7] [8] and fusing the results with a state-of-the-art system. Others try to use per frame sampled short-term prosodic values like intonation and power curves [9] [10]. This last approach is very appealing, as the new features could be easily combined with the traditional cepstral coefficients.

This work focuses on determining whether simple short-term prosodic information is useful for improving current state-of-the-art ASV. In order to see this, a new ASV system is presented, which uses both spectral and prosodic features. The paper is organized as follows: First the constructed verification system is explained. Then, a description of the database used during the experiments is given, followed by a description of the experiments and their results. Finally, these results are commented.

2. DESCRIPTION OF THE VERIFICATION SYSTEM

2.1. Baseline system

The baseline system consists of a traditional GMM-UBM system [11] with Mel Frequency Cepstral Coefficients (MFCC) parameterization. More precisely,

This work has been partially funded by the Spanish Ministry of Science and Technology (TIC2003-08382-C05-03).

an 18 MFCC feature vector was extracted every 10 milliseconds. These vectors were augmented with the first and second order derivatives. Cepstral Mean Subtraction (CMS) was applied in order to reduce the channel effects [12].

Speaker models were trained by Maximum A Posteriori (MAP) adaptation of the previously trained Universal Background Model (UBM) [3]. Only the means were adapted, leaving the variances and weights unaltered. Last, as usual in this kind of systems, UBM and HNorm score normalization [13] were performed during the tests.

2.2. Prosodic system

In the prosody based system, energy and intonation related information was used. Separate models were created for prosody modelling in voiced and unvoiced regions, in order to handle the discontinuities in the intonation curve. The signal's power was estimated every 10 milliseconds using Hamming windowing of 30 milliseconds length. F0 was also estimated every 10 milliseconds, using a method based on cepstrum transformation and Viterbi algorithm. This method not only gives the estimated value of F0, but also decides whether a frame is voiced or not. Once the power and intonation curves were estimated, their first and second order derivatives were calculated, in order to take into account the dynamics of the features.

The voiced and unvoiced frames were separated to get two feature vector streams. Voiced frames were parameterized with five features (instantaneous F0, its first and second derivative, and the first and second derivative of the power), whereas unvoiced frames were parameterized with only two (the first and second derivatives of the power). The instantaneous power was discarded in both cases, as this value is more related to the channel gain than to the speaker's identity.

Using these two vector streams, two different models were trained for each speaker, using a traditional GMM-UBM scheme. First, two UBM models were developed, one for voiced and another one for unvoiced frames. Then, the speaker models were created from these by means of MAP adaptation.

During the test phase, and for each recording, two different scores were calculated, one for the voiced and another one for the unvoiced frame streams. UBM score normalization was applied to these scores before fusing them with the product rule. That is, the final score for the prosodic information is the product of the scores of the voiced and the unvoiced stream scores.

2.3. System fusion

In order to combine the results of both classifiers, a late fusion scheme was chosen: first the scores using both traditional and prosody based systems are calculated, and then, a final score is obtained combining these scores.

For this purpose Linear Discriminant Algorithm (LDA) [14] was selected. LDA is capable of finding the linear combination of scores that best separates user and impostor scores, but it must be trained over a validation set of speakers, before applying it to the final tests.

3. DATABASE DESCRIPTION

The AHUMADA database [15] was used for the experiments carried out in this work. This database consists of voice recordings of 103 male Spanish speakers, and was specifically recorded for the development of automatic speaker recognition systems. In fact, it was used during the NIST speaker recognition evaluation campaigns of 2000 and 2001 [2], together with an extension aimed to include feminine speakers.

Although the complete database contains both high quality microphone recordings and telephone speech recordings, only the later ones have been used for the experiments. This allows capturing the effects that channel distortion has on the system's behaviour. These telephonic recordings were carried out along three different sessions, and in each of them, the speakers used a different telephone handset.

- The first session (named T1), was recorded through an internal-routing call with a flat frequency response, so there is no spectral distortion in the signals.
- In the second session (T2), the recording was done through the speaker's home telephone, so the handset type remains unknown.
- In the third session (T3) each of the speakers used one of the nine handsets that were available in the recording laboratory. So that, for this session and for each speaker, the handset type that was used (carbon button or electret) is known.

Among the items that were recorded in each session, two were selected for the experiments: The read common text (the same text for all speakers and sessions) and the read specific text (a different text for each of the speakers and sessions). From now on, these items will be named C (for Common) and S (for Specific) respectively. All these recordings were sampled at 8 kHz, 16 bit per sample. The mean length of the selected items was about 65 seconds.

To sum up, three telephonic sessions of 103 male speakers, with two items per session were available for the experiments. This sub-corpus was divided as follows: 51 speakers were reserved to train the UBM models, 26 were used for the validation tests, and the rest were kept as users of the system. Every user was treated as an impostor for the rest of the users, which is equivalent to having 25 impostors. The speakers were randomly designated to one of the groups, in order not to bias the result.

# mix.	*2*	*4*	*8*	*16*	*32*	*64*	*128*	*256*	*512*	*1024*
Baseline	29.13	20.29	17.18	13.25	10.36	8.12	6.75	5.71	5.21	3.85
Voiced	35.47	33.21	29.88	28.96	27.78	25.32	23.93	24.03	24.74	23.93
Unvoiced	46.53	49.45	47.86	45.86	44.44	43.16	42.74	43.18	43.59	44.44
V+UV	35.27	32.48	29.49	28.77	27.33	25.64	23.93	24.89	24.53	24.14

Table 1. *%EER values for the baseline and prosody based systems.*

Speaker model training was done using tasks C and S from session T1. As this session was recorded without spectral distortion, the resulting models will be channel independent. Tasks C and S from session T3 were used for the development tests, because knowing which handset type was used makes it possible to determine, the HNorm coefficients for handset normalization. Finally, task S from session T2 was reserved for the final tests. In this way, these final tests were carried out with a handset that was unknown for the system, as it was not used before, neither in the training nor in the development phase. Furthermore, using only task S, also the contents of the recordings were new for the system.

For the training, the whole recordings were used. This gives about 130 seconds of training material for each speaker. For the tests, nine speech segments of 10 seconds each were extracted from each item, allowing a 50% overlap between two consecutive segments. So, for each speaker 9 validation test segments were available. This means that each user was tested against 25x9=225 impostor segments.

All this makes the design of these experiments very realistic, as in an environment in which the users make the recordings for training and development in the laboratory, but they try to gain access from their own home or office, using their own phone.

4. RESULTS

In a GMM system, the number of Gaussian mixtures is critical for the system's accuracy. In order to estimate the best number of mixtures for the models, different order GMMs were developed, from 2 up to 1024 mixtures, both for MFCC and prosody based systems. The final mixture number was selected in order to minimize the Equal Error Rate (EER) between False Rejection (FR) and False Acceptance (FA) rates. Table 1 shows the ERR curve for the trained systems. As expected, the baseline system achieves a better performance than prosody alone. It can also be seen that energy information of unvoiced frames does not help to the final result of the prosody based verification system.

According to the achieved EER values, 1024 and 128 mixture GMMs were selected for the baseline and prosody based systems respectively. After applying LDA and fusing the scores, a final EER of 3.84% was achieved. Figure 1 shows the DET curves [16] of the systems before and after applying the fusion. The DET curves for the baseline and the combined systems are very similar. In fact, both systems have the same EER.

5. CONCLUSIONS

As a result of these experiments, it can be seen that adding simple intonation and energy related short-term features does not improve the state-of-the-art results. Due to the great improvement achieved in the last few years in ASV (like various handset and likelihood normalization techniques), current spectral envelope based systems achieve a performance much greater than the prosody based ones (3.85% against 23.93% EER in these experiments).

This does not mean that prosody is not useful for ASV, as current researches have achieved some improvement using it [9] [10], but that prosody based ASV is far behind spectra based one. For example, just as HNorm came out to deal with handset variability in traditional systems, there is a need to solve the inter-session variability in prosody. There are still very interesting research areas to explore in this field.

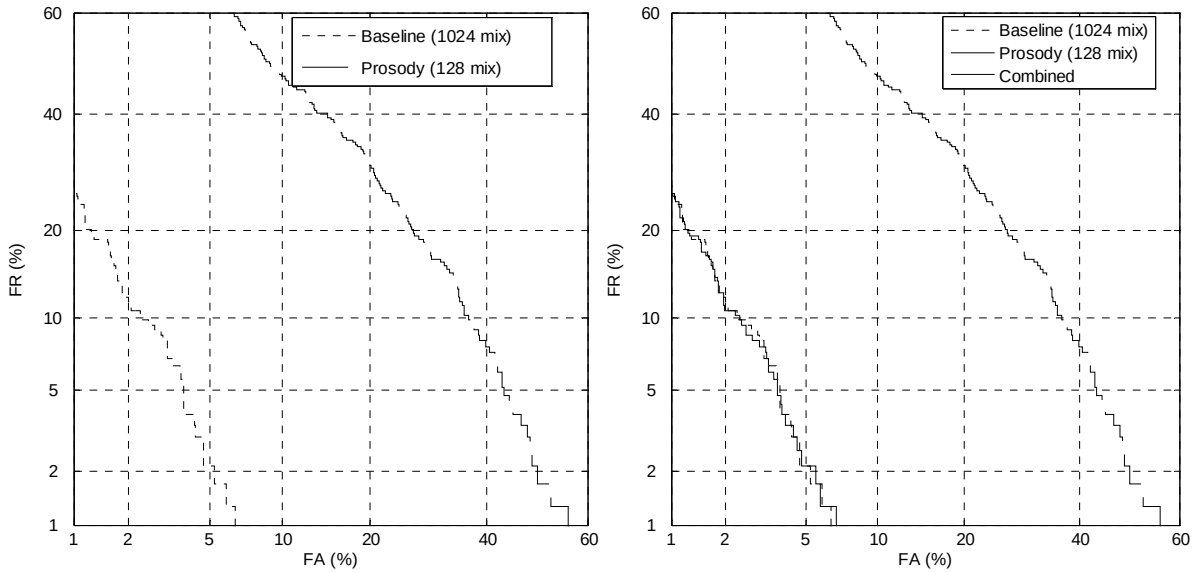

Figure 1. *DET curves for the baseline and prosody based systems (left) and the final combined system (right).*

6. REFERENCES

[1] FVC2006 web site: http://bias.csr.unibo.it/fvc2006/

[2] NIST Speaker Recognition Evaluations' web site: http://www.nist.gov/speech/tests/spk/

[3] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker Verification Using Adapted Gaussian Mixture Models," Digital Signal Processing, vol. 10, pp. 19-41, 2000.

[4] M. A. Przybocki and A. F. Martin, "NIST Speaker Recognition Evaluation Chronicles," presented at Odyssey, Toledo, España, 2004.

[5] A. F. Martin and M. A. Przybocki, "The NIST Speaker Recognition Evaluations: 1996-2001," presented at Speaker Odyssey 2001, Creta, Grecia, 2001.

[6] M. J. Carey, E. S. Parris, H. Lloyd-Thomas, and S. Bennett, "Robust Prosodic Features for Speaker Identification," presented at ICSLP, Philadelphia, EEUU, 1996.

[7] A. G. Adami, R. Michaescu, D. A. Reynolds, and J. J. Godfrey, "Modeling Prosodic Dynamics for Speaker Recognition," presented at ICASSP, Hong Kong, 2003.

[8] B. Peskin, J. Navratil, J. Abramson, D. Jones, D. Klusacek, D. A. Reynolds, and B. Xiang, "Using Prosodic and Conversational Features for High Performance Speaker Recognition," presented at ICASSP, Hong Kong, 2003.

[9] D. A. Reynolds, W. Andrews, J. P. Campbell, J. Navratil, B. Peskin, A. G. Adami, Q. Jin, D. Klusacek, J. Abramson, R. Michaescu, J. J. Godfrey, J. Douglas, and B. Xiang, "SuperSID project final report," SuperSID project (http://www.clsp.jhu.edu/ws2002/groups/supersid/) 2002.

[10] M. Arcienega and A. Drygajlo, "A Bayesian Network Approach for Combining Pitch and Spectral Envelope Features for Speaker Verification," presented at COST275 workshop, Roma, Italia, 2002.

[11] F. Bimbot, J. F. Bonastre, C. Fredouille, G. Gravier, I. Magrin-Chagnolleau, S. Meignier, T. Merlin, J. Ortega-García, D. Petrovska-Delacrétaz, and D. A. Reynolds, "A Tutorial on Text-Independent Speaker Verification," EURASIP Journal on Applied Signal Processing, vol. 4, pp. 430-451, 2004.

[12] C. Barras and J. L. Gauvain, "Feature And Score Normalization for Speaker Verification of Cellular Data," presented at ICASSP, Hong Kong, 2003.

[13] R. Auckenthaler, M. J. Carey, and H. Lloyd-Thomas, "Score Normalization for Text-Independent Speaker Verification Systems," Digital Signal Processing, vol. 10, pp. 42-54, 2000.

[14] J. B. Kennedy and A. M. Neville, *Basic Statistical Methods for Engineers and Scientists*, Harper & Row, New York, 1986.

[15] J. Ortega-García, J. González-Rodríguez, and V. Marrero-Aguiar, "AHUMADA: A Large Speech Corpus in Spanish for Speaker Characterization and Identification," Speech Communication, vol. 31, pp. 255-264, 2000.

[16] A. F. Martin, G. R. Doddington, T. Kamm, M. Ordowski, and M. A. Przybocki, "The DET Curve in Assessment of Detection Task Performance," presented at Eurospeech, Rhodes, Grecia, 1997.