Channel / Handset Mismatch Evaluation in a Biometric Speaker Verification Using Shifted Delta Cepstral Features

José R. Calvo, Rafael Fernández, and Gabriel Hernández

Advanced Technologies Application Center, CENATAV, Cuba
{jcalvo,rfernandez,gsierra}@cenatav.co.cu

Abstract. This paper examines the application of Shifted Delta Cepstral (SDC) features in biometric speaker verification and evaluates its robustness to channel/handset mismatch due by telephone handset variability. SDC features were reported to produce superior performance to delta features in cepstral feature based Language Identification systems. The result of the experiment reflects superior performance of SDC features regarding to delta features in biometric speaker verification using speech samples from Ahumada Spanish database.

Keywords: biometrics, speaker verification, cepstral features, shifted delta cepstral features, channel mismatch.

1 Introduction

Existing methods of user authentication can be grouped into three classes: possessions (something that you have: a key, an identification card, etc); knowledge (something that you know: a password, a PIN, etc) and biometrics [1]. Biometrics is the science of identifying or verifying the identity of a person based on physiological characteristics (something that you are: fingerprints or face) or behavioural characteristics dependent on physical characteristics (something that you produce: handwritten signature or speech).

Early user authentication was based on possessions and knowledge, but problems associated with these methods, restrict their use. The most important drawbacks of these methods are: possessions can be lost, stolen, shared or easily duplicated; knowledge can be shared, easy to guess, forgotten, and both, knowledge and possessions can be shared or stolen [1]. Consequently it is easy to deny that a given person carried out an action, because only the possessions or knowledge are checked, and these are only loosely coupled to the person's identity. Biometrics provides a solution to these problems by truly verifying the identity of the individual.

As a biometric user authentication method, speech is a behavioural characteristic that is not considered threatening or intrusive by users to provide. The goal of speaker recognition is to extract, characterize, and recognize the information in the speech signal conveying speaker identity [2]. Telephony is the main modality of biometric speaker recognition, since it is a domain with ubiquitous existing hardware and doesn't need for special transducers to be installed.

L. Rueda, D. Mery, and J. Kittler (Eds.): CIARP 2007, LNCS 4756, pp. 96–105, 2007.

Current automatic speaker recognition systems face significant challenges caused by adverse acoustic conditions as telephone band limitation and channel and handset variability. Degradation in the performance of speaker recognition systems due to channel mismatch has been one of the main challenges to actual deployment of speaker recognition technologies. Several techniques have been proposed to address this problem, new speech features that are less sensitive to channel effects can be extracted [3], the effect of mismatches can be reduced via cepstral normalization [4, 5], the speaker models can be transformed to compensate for the mismatches [6, 7], and rescoring techniques can be used to normalize the speaker scores and reduce the channel and handset effects [8].

This paper introduces the application of a new set of dynamic cepstral features in speaker recognition: Shifted Delta Cepstral (SDC) features, and evaluates its performance in front of channel/handset mismatch, typical in remote applications. SDC features were recently reported to produce superior performance to delta features in cepstral feature based Language identification [9, 10].

SDC features are obtained by concatenating the delta-cepstral computed across multiple frames of speech. As a combination of dynamic cepstral features, SDC features contain useful information about speaker identity.

Nevertheless, in our knowledge, this is the first attempt on using SDC features for speaker recognition. This evaluation was performed using telephone speech samples of Ahumada Spanish database [11].

2 Biometric Speaker Verification

Voice is a combination of physiological and behavioral characteristics. The features of an individual's voice are based on invariant physiological characteristics, as the shape and size of the vocal and nasal tract, mouth and lips, used in the synthesis of the sound. Nevertheless, this technology is usually classified as a behavioural too, because the way the individual speaks, their attitude and their cultural background strongly influences the resulting speech signal. This behavioral characteristics of a person's speech (and some physiological, too) changes over time due to age, health conditions, emotional state, environmental reasons, etc.

Biometric application of speaker recognition is identified as speaker verification because a user claims to be a client, and the system verifies this claim. Many applications of speaker verification systems are accessed remotely by users and the channel involved in the communication is the telephone. Because the handset and the line can vary from call to call, there is often an acoustic mismatch between the speech collected to train the speaker models and the speech produced by the speakers at run time or during testing. Such mismatches are known to severely affect the performance of the system. However, in a remote banking application, the voice-based technique combined with other user's authentication method, may be preferred since it can be integrated without additional effort, into the existing telephone system.

Speaker verification systems are categorized depending on the freedom in what is spoken; this taxonomy based on increasingly complex tasks also corresponds to the sophistication of algorithms used and the progress in the art over time [1]:

Fixed text: The speaker says a predetermined word or phrase which was recorded at enrolment. The word may be secret, so it acts as a password, but once recorded a replay attack is easy, and re-enrolment is necessary to change the password.

Text prompted: The speaker is prompted by the system to say a specific expression. The system matches the utterance with known text to determine the user. For this, enrolment is usually longer, but the prompted text can be changed at will. Expression as digit strings are more vulnerable than phrases, to splicing-based replay attacks.

Text independent: The system processes any utterance of the speaker. Here the speech can be task-oriented, so it is hard to acquire speech that also accomplishes the impostor's goal.

Combined with utterance verification [2]**:** The system presents to the user, a series of randomized phrases to repeat, and verifies not only the voice matches but also the required phrases match. Additionally, it is possible to use forms of automatic knowledge verification where a person is verified by comparing the content of his/her spoken utterance against the stored information in his/her personal profile.

This paper evaluates the performance of SDC features as a new set of dynamic features for speaker recognition, in a remote speaker verification system using text prompted task using short phrases.

3 Shifted Delta Cepstral Features

First proposed by Bielefeld [12], features called Shifted Delta Cepstral (SDC) are obtained by concatenating the delta-cepstral computed across multiple frames of speech information, spanning multiple frames into the feature vector. Recently, the proposal of using SDC features of a speech signal for language identification with GMM [13] and SVM [14] classifiers, has produced promising results. In our knowledge, this is the first attempt to using SDC for speaker recognition.

Cepstral features contain information about speech formants structure, and delta-cepstral about its dynamics. SDC features evaluate speech spectral dynamics better, because can reflect the movement and position of vocal and nasal articulators if its time interval of analysis is adjusted to include spectral transitions between phonemes and syllables. In each cepstral frame, SDC computation obtains the dynamic of the articulatory movement in next frames, as a pseudo-prosodic feature vector [10] computed without having to explicitly find or model the prosodic structure of the speech signal. Is known that the prosodic structure of the speech conveys important information about the identity of the speaker [15].

The computation of SDC features is a relatively simple procedure [16] and is illustrated in Fig. 1. First, a cepstral feature vector is computed in each frame. A shifting delta operation is applied to frame based cepstral feature vectors in order to create the new combined feature vectors for each frame.

The SDC features are specified by a set of 4 parameters, (*N, d, P, k*) where:

- *N:* number of c cepstral coefficients in each cepstral vector.
- *d:* time advance and delay for the delta computation.

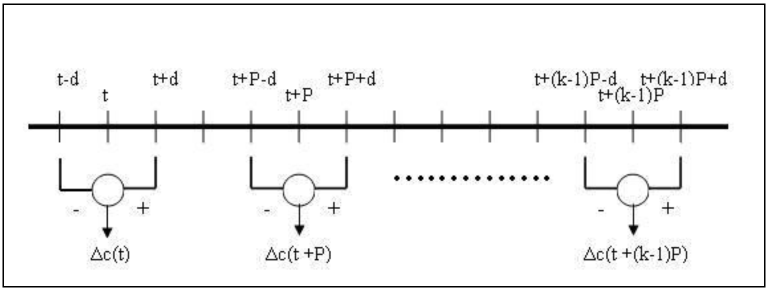

Fig. 1. Computation of SDC feature vector for each cepstral coefficient

- *P:* time shift between consecutive blocks.
- *k:* number of blocks whose delta coefficients are concatenated to form the SDC vector

For the case shown in Fig 1 the final SDC vector at frame time *t* is given by the concatenation from $i = 0$ to $k\text{-}1$ of all the $\Delta c\ (t + iP)$, where:

$$\Delta c(t+iP)=c(t+iP+d)-c(t+iP-d) \tag{1}$$

Accordingly, kN parameters are used for each SDC feature as compared with 2N for conventional cepstral and delta-cepstral feature vectors. In language identification applications, SDC features substitute cepstral and delta-cepstral features, using different combinations of (N, d, P, k).

More recently, a modified version of SDC was reported to have even higher performance in LID [9], calculated using a recurrent expression:

$$\Delta c(t+iP)=\frac{\sum_{d=-D}^{D} dc(t+iP+d)}{\sum_{d=-D}^{D} d^2} \tag{2}$$

4 Front End Processing

Cepstral coefficients derived from a Mel-frequency filter bank (MFCC) have been used to represent the short time speech spectra. All speech material used for training and testing is pre-emphasized with a factor of 0.97, and an energy based silence removal scheme is used. A Hamming window with 30ms window length and 30% shift is applied to each frame and a short time spectrum is obtained applying a FFT. The magnitude spectrum is processed using a 30 Mel-spaced filter bank, the log-energy filter outputs are then cosine transformed to obtain 12 Mel-frequency cepstral coefficients, the zero cepstral coefficient is not used. Therefore, each window of signal frame is represented by a 12-dimensional MFCC features vector.

In order to reduce the influence of mismatch between training and testing acoustic conditions, a robust feature normalization method for reducing noise and/or channel effects has been proposed, the Cepstral Mean and Variance Normalization (CMVN) [16]. Assuming Gaussian distributions, CMVN normalizes each component of the feature vector according to the expression:

$$\hat{c}_i[n] = \frac{c_i[n] - \mu_i}{\sigma_i} \tag{3}$$

where $c_i[n]$ and $\hat{c}_i[n]$ are the i-th component of the feature vectors at time frame n before and after normalization, respectively, and μ_i and σ_i are the mean and variance estimates of the sequence $c_i[n]$.

Delta-cepstral features are obtained for each MFCC features vector, using *d=2* as time advance and delay for the delta computation, at last, and using equation 2, SDC features are obtained.

Three set of features are used in each one of the experiments:

1. 12 MFCC + 12 delta , dimension 24 (baseline) : MFCC + D
2. 12 MFCC + SDC (12,2,2,2), dimension 36: MFCC + SDC
3. 12 SDC(12,2,2,2), dimension 24: SDC

5 Database and Experiments

Ahumada [11] is a speech database of 103 Spanish male speakers, designed and acquired under controlled conditions for speaker characterization and identification. Each speaker in the database expresses six types of utterances in seven microphone sessions and three telephone sessions, with a time interval between them.

In order to evaluate the performance of SDC features in front to handset and channel mismatch in a remote biometric speaker verification using text prompted phrases, ten phonologically and syllabically balanced phrases in the three telephone sessions of Ahumada were used, the ten phrases are the same for each one of the 103 speakers. The performance of the verification is evaluated using a 64 mixtures GMM/UBM classifier, trained and tested with a subset of 50 speakers of the database; other subset of 50 speakers is used to train the 256 mixtures UBM.

In our approach, the behaviour of a text prompted biometric speaker verification is simulated, so the system is trained with ten phrases of each one of 50 speakers in session T1 and tested with each one of the phrases of the same speakers in session T2 and T3. All 50 speakers were used as targets for their corresponding models and as impostors for the rest of models, so we obtain 500 target and 4500 impostors in each test.

In each telephone sessions, conventional telephone line was used. In session T1, every speaker was calling from the same telephone, in an internal-routing call. In session T2, all speakers were requested to make a call from their own home telephone, trying to search a quiet environment, so the channel and handset characteristics are unknown. In session T3, a local call was made from a quiet room, using 9 randomly selected standard handsets, for each handset, three characteristics are known: microphone sensibility and frequency response, and the ranges of signal to noise ratio in its associated channel.

Each speaker in session T3 uses one of the 9 handset, then the speakers can be grouped in two classes, for each one of the three measured characteristics:

- Low sensibility (< 1 mV/P) and high sensibility (> 2.5 mV/P) of the microphone.
- Low attenuation level (< 20 dB) and high attenuation level (> 35 dB) of the microphone band pass frequency response.
- Low and high signal to noise ratio mean (threshold: 35 dB) in the channel.

The experiments are organized in the following manner:

1. Evaluation of channel mismatch in uncontrolled conditions: trained with session T1 and tested with session T2
2. Evaluation of channel mismatch due to handset sensibility: trained with speakers in session T1 and tested with speakers in session T3, grouped in two classes, low sensibility (24 speakers) and high sensibility (26 speakers).
3. Evaluation of channel mismatch due to handset frequency response: trained with speakers in session T1 and tested with speakers in session T3, grouped in two classes, low attenuation level (30 speakers)and high attenuation level (20 speakers).
4. Evaluation of channel mismatch due to signal to noise ratio in the channel: trained with speakers in session T1 and tested with speakers in session T3, grouped in two classes, low (19 speakers) and high (31 speakers) signal to noise ratio mean.

6 Results

Evaluation of the results was performed using detection error tradeoff (DET) plot [17].Two indicators are used to evaluate the performance: Equal error rate (EER) and minimum of Detection Cost Function (DCF), defined as:

$$\mathrm{DCF}= (\mathbf{C}_{\mathbf{FR}} * \mathbf{P}_{\mathbf{FR}} * \mathbf{P}_{\mathbf{Target}}) + (\mathbf{C}_{\mathbf{FA}} * \mathbf{P}_{\mathbf{FA}} * \mathbf{P}_{\mathbf{NonTarget}}) \quad (4)$$

Where

$\mathbf{C}_{\mathbf{FR}}$ (cost of a missed detection) = 10
$\mathbf{C}_{\mathbf{FA}}$ (cost of a false alarm) = 1
$\mathbf{P}_{\mathbf{Target}}$ (a priori probability of a target speaker) = 0.01
$\mathbf{P}_{\mathbf{NonTarget}}$ (a priori probability of a non-target speaker) = 0.99
$\mathbf{P}_{\mathbf{FR}}$ (Miss probability)
$\mathbf{P}_{\mathbf{FA}}$ (False alarm probability)

The results of the four experiments are reflected in DET plots in figures 2 to 5 and Tables 1 to 4 with values of indicators EER and DCF.

DET plot of experiment 1 reflects a similar behaviour of SDC and MFCC features in front of channel mismatch where the channel and handset characteristics are unknown. Table 1 shows that MFCC + SDC features have better performance that MFCC +D features (better EER and DCF).

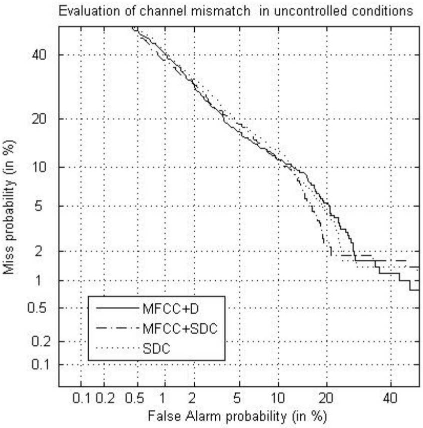

Fig. 2. Experiment 1: T1 train, T2 test

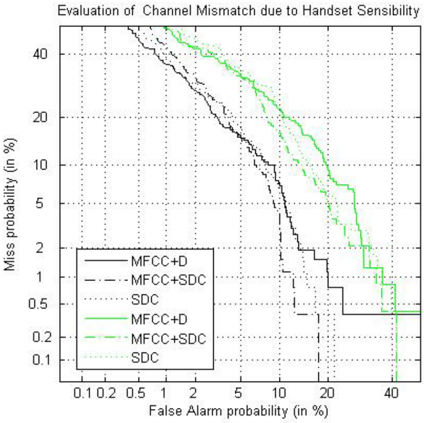

Fig. 3. Experiment 2: T1 train, T3 test black : high sensibility , green: low sensibility

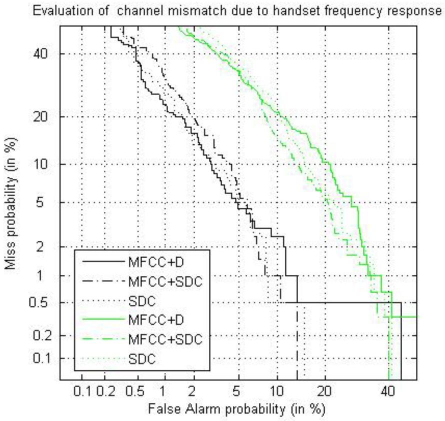

Fig. 4. Experiment 3: T1 train, T3 test black : low attenuation, green: high attenuation

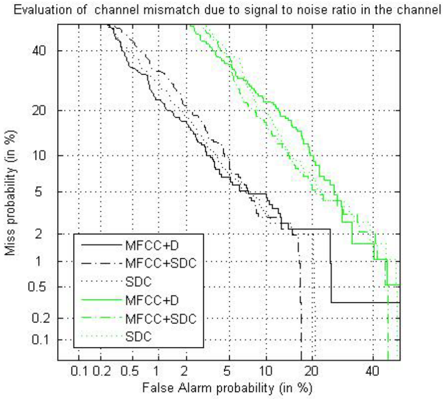

Fig. 5. Experiment 4: T1 train, T3 test black : high s/n, green: low s/n

Table 1. Experiment 1: channel mismatch in uncontrolled conditions

Features	EER	DCF
MFCC +D	0.107	0.048
MFCC+SDC	0.102	0.047
SDC	0.111	0.051

DET plot of experiment 2 reflects a better behaviour of both sets of SDC features compared to MFCC features in front of mismatch due to handset sensibility . Table 2 shows that both sets of SDC features have lower EER and similar DCF that MFCC +D features in both sensibility conditions.

Table 2. Experiment 2: channel mismatch due to handset sensibility

	Low sensibility		High sensibility	
Features	EER	DCF	EER	DCF
MFCC +D	0.154	0.057	0.091	0.045
MFCC+SDC	0.119	0.057	0.080	0.049
SDC	0.132	0.061	0.084	0.046

Table 3. Experiment 3: channel mismatch due to handset frequency response

	High attenuation		Low attenuation	
Features	EER	DCF	EER	DCF
MFCC +D	0.154	0.062	0.049	0.031
MFCC+SDC	0.12	0.064	0.055	0.037
SDC	0.133	0.067	0.05	0.032

DET plot of experiment 3 reflects a better behaviour of both sets of SDC features compared to MFCC features in front of high attenuation of handset frequency response. Table 3 shows that both sets of SDC features have lower EER and similar DCF that MFCC +D features in this condition.

Table 4. Experiment 4: channel mismatch due to signal to noise ratio in the channel

	Low s/n		High s/n	
Features	EER	DCF	EER	DCF
MFCC +D	0.157	0.070	0.058	0.032
MFCC+SDC	0.121	0.072	0.063	0.039
SDC	0.127	0.074	0.061	0.035

DET plot of experiment 4 reflects a better behaviour of both sets of SDC features compared to MFCC features in front of low signal to noise ratio in the channel. Table 3 shows that both sets of SDC features have lower EER and similar DCF that MFCC +D features in this condition.

Results of experiments 2, 3 and 4 reflect a better performance of both sets of SDC features in front of the worst mismatch condition: low handset sensibility, high attenuation in handset frequency response and low signal to noise ratio in the handset associated channel. Table 5 reflects the relative reduction in % of EER, in each experiment for both sets of SDC features respect to MFCC features.

Table 5. Reduction in % of EER for both sets of SDC features respect to MFCC features

Mismatch condition	MFCC + SDC	SDC
low handset sensibility	22	14
high handset attenuation	22	13
low s/n ratio in channel	23	19

7 Conclusions and Future Work

The result of the experiments reflect a superior performance of SDC features respect to MFCC + delta features in speaker verification using speech samples from telephone sessions of Ahumada Spanish database.

- Test in uncontrolled conditions (experiment 1) reflects similar behavior of SDC and MFCC features.
- Tests under controlled conditions (experiments 2, 3 and 4) reflect a better behaviour of SDC respect to MFCC features in front of worst mismatch conditions.
- In these experiments, the EER reduction due to utilization of SDC features instead of MFCC features is superior to 22% using MFCC+SDC, and superior to 13% using SDC alone.
- Test under controlled conditions (experiment 2,3 and 4) reflect a similar behavior of SDC respect to MFCC features in front to better mismatch conditions, in experiment 2, SDC features have a better behavior than MFCC features in both mismatch conditions.

Shifted Delta Cepstral features must be considered as a new alternative of cepstral features, in order to reduce the effects of channel/handset mismatch in remote speaker verification performance. SDC features appended to MFCC features show the best results, but SDC features instead of MFCC +delta features show a good result too, maintaining the same feature dimensionality (24 dimensions).

Future work will be in the direction of evaluate the influence of SDC parameters d and P. SDC features must be assumed as a pseudo-prosodic vector, and these parameters are related with its time-dynamic behaviour. Also, H-Norm score normalization must be applied.

References

1. Ratha, N.K., Senior, A., Bolle, R.M.: Automated Biometrics. In: Singh, S., Murshed, N., Kropatsch, W.G. (eds.) ICAPR 2001. LNCS, vol. 2013, pp. 445–474. Springer, Heidelberg (2001)
2. Ortega-Garcia, J., Bigun, J., Reynolds, D., Gonzalez-Rodriguez, J.: Authentication gets personal with biometrics. IEEE Signal Processing Magazine 50–62 (2004)
3. Heck, L.P., Konig, Y., Sonmez, M.K., Weintraub, M.: Robustness to telephone handset distortion in speaker recognition by discriminative feature design. Speech Communication 31, 181–192 (2000)
4. Mammone, R., Zhang, X., Ramachandran, R.: Robust speaker recognition. IEEE Signal Processing Magazine 58–71 (1996)
5. Rahim, M.G., Juang, B.H.: Signal Bias Removal by Maximum Likelihood Estimation for Robust Telephone Speech Recognition. IEEE Trans. On Speech and Audio Processing 4(1), 19–30 (1996)
6. Yiu, K.K., Mak, M.W., Kung, S.Y.: Environment Adaptation for Robust Speaker Verification. In: Eurospeech 2003, Geneva, pp. 2973–2976 (2003)
7. Teunen, R., Shahshahani, B., Heck, L.P.: A model based transformational approach to robust speaker recognition. In: Proc. ICSLP (2000)

8. Reynolds, D.A: Comparison of background normalization methods for text-independent speaker verification. Proceedings European Conf. on Speech Communication and Technology. Eurospeech (1997)
9. Allen, F.: Automatic Language Identification. PhD Thesis, University of New South Wales, Sydney, Australia (2005)
10. Lareau, J.: Application of Shifted Delta Cepstral Features for GMM Language Identification. MsC Thesis, Rochester Institute of Technology, USA (2006)
11. Javier, O.-G., Joaquin, G.-R., Victoria, M.-A.: AHUMADA A Large Speech Corpus in Spanish for Speaker Characterization and Identification. Speech Communication (31), 255–264 (2000)
12. Bielefeld, B.: Language identification using shifted delta cepstrum. In: Proc. Fourteenth Annual Speech Research Symposium (1994)
13. Torres-Carrasquillo, P.A., Singer, E., Kohler, M.A., Greene, R.J., Reynolds, D.A., Deller Jr., J.R.: Approaches to language identification using Gaussian Mixture Models and shifted delta cepstral features. In: Proc. ICSLP, pp. 89–92 (2002)
14. Singer, E., Torres-Carrasquillo, P.A., Gleason, T.P., Campbell, W.M., Reynolds, D.A.: Acoustic, Phonetic, and Discriminative Approaches to Automatic Language Recognition. In: Proc. Eurospeech 2003, pp. 1345–1348 (2003)
15. Reynolds, D., Andrews, W., Campbell, J., Navrátil, J., Peskin, B., Adami, A., Jin, Q., Klusáček, D., Abramson, J., Mihaescu, R., Godfrey, J., Jones, D., Xiang, B.: Supersid final report: exploiting high-level information for high-performance speaker recognition. Tech. Rep. Workshop, The Centre for Language and Speech Processing (2002)
16. de Wet, F.: Additive Background Noise as a Source of non-Linear Mismatch in the Cepstral and Log-Energy Domain. Computer Speech and Language 19, 31–54 (2005)
17. Martin, A., et al.: The DET curve assessment of detection task performance. Proc. of EuroSpeech 4, 1895–1898 (1997)