INTERSPEECH 2009 BRIGHTON

Selection of the Best Set of Shifted Delta Cepstral Features in Speaker Verification Using Mutual Information.

José R. Calvo, Rafael Fernández, Gabriel Hernández

Advanced Technologies Application Centre, CENATAV
Calle 7ª #21812, Playa, Ciudad Habana, Cuba
{jcalvo,rfernandez,gsierra}@cenatav.co.cu

10.21437/Interspeech.2009-393

Abstract

Shifted delta cepstral (SDC) features, obtained by concatenating delta cepstral features across multiples speech frames, were recently reported to produce superior performance to delta cepstral features in language and speaker recognition systems. In this paper, the use of SDC features in a speaker verification experiment is reported. Mutual information between SDC features and identity of a speaker is used to select the best set of SDC parameters. The experiment evaluates robustness of the best SDC features due to channel and handset mismatch in speaker verification. The result reflects an EER relative reduction until 19% in a speaker verification experiment.

Index Terms: speaker verification, shifted delta cepstral, mutual information

1. Introduction

As a biometric user authentication method, voice is a behavioral characteristic that is not considered threatening or intrusive by users. The goal of speaker recognition is to extract, characterize, and recognize the information in the voice signal conveying speaker identity [1].

Telephony is the main modality of biometric speaker recognition, since it is a domain with ubiquitous existing hardware where no other biometric can be used and does not need for special transducers to be installed. Although adverse acoustic conditions, telephone band limitation and channel and handset variability represent significant challenges.

This work uses a new dynamic cepstral features in speaker recognition -Shifted Delta Cepstral (SDC) features- and evaluates its robustness in front of channel and handset mismatch, typical in remote speaker verification. SDC features, obtained by concatenating delta cepstral features computed across multiple frames of speech, were first reported by Torres-Carrasquillo *et al.* [2] to produce superior performance to delta features in language recognition. More recently, the authors [3] report a superior performance of one SDC feature set, respect to delta features in speaker verification under mismatch conditions.

Mutual Information (MI) has been recently used in speaker verification as a feature selection technique [4, 5]. In this paper, the authors propose its use to select the better set of SDC parameters, evaluating the MI between the SDC features and the identity of speakers, and selecting those SDC features with better MI, to explore their robustness in front of channel and handset mismatch in a text prompted speaker verification experiment, using telephone balanced phrases from NIST 2001 Ahumada database [6].

The rest of the paper is organized as follows. In Section 2, SDC features are described, experiment front-end and procedure to obtain SDC features are explained; at last, used Database is described. Section 3 describes the use of MI to evaluate the quantity of information that SDC features conveys about the speaker, and reflects the MI between SDC feature sets and speaker identity. In Section 4, a speaker verification experiment is performed with selected SDC feature sets. Finally, Section 5 concludes this work and gives a future research direction.

2. Shifted delta cepstral features

SDC is a long term dynamic feature obtained by concatenating several Δ cepstral features across multiple frames of speech information, into one vector. The SDC feature is specified by a set of 4 parameters (*N, D, P, k*):

- *N:* number of *c* coefficients in each cepstral vector.
- *D:* time advance and delay for the delta computation.
- *P:* time shift between consecutive frames.
- *k:* number of frames whose Δ features are concatenated.

The computation of SDC feature is a simple procedure, first, an N-dimension cepstral vector is computed in each speech frame *t,* and then each *c* coefficient is differenced using spaced $t \pm D$ frames to obtain the Δ feature:

$$\Delta c(t) = \frac{\sum_{d=-D}^{D} dh_d c(t+d)}{\sum_{d=-D}^{D} h_d d^2} \qquad (1)$$

Then, for each *c* coefficient, *k* different Δ features, spaced *P* frames, are stacked to form a SDC (*c, D, P, k*) vector, for each frame *t* as is illustrated in Figure 1.

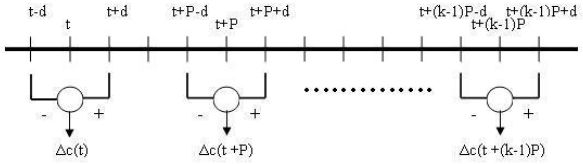

Figure 1 *Computation of SDC feature vector for each cepstral coefficient*

So, the SDC feature for each *c* coefficient at frame time *t,* does not require extra computational cost and *is* given by the concatenation, from *i* = 0 to k-*1* blocks of all the Δ*c*(*t*+*iP*):

$$\text{SDC}(c, D, P, k) = \Delta c(t),\ \Delta c(t+P),\ \ldots\ldots,\ \Delta c(t+(k-1)P) \qquad (2)$$

2.1 Combinations of SDC parameters

SDC computation is controlled by four parameters *(N, D, P, k),* each one has its influence in the vector dimensionality, computational cost and pseudo-prosodic behavior of the SDC:

- *N*: establish the number of c cepstral coefficients in each cepstral vector; in conjunction with *k*, determine the vector dimensionality.
- *D:* time advance and delay for the delta computation in the equation 1, determine the time width of the generalized spectral slope and the computational cost.
- *P*: time shift between consecutive frames, establish, in conjunction with *D*, the overlapped number of frames of SDC vector , as *D* increases, is necessary a greater value of *P* to avoid the overlap of the frames. It doesn't affect the computational cost or the vector dimensionality.
- *k*: number of frames whose Δ features are concatenated to form the SDC vector; determine the pseudo-prosodic behavior of the SDC feature, and in conjunction con *N*, the dimensionality of the vector. It doesn't affect the computational cost.

2.2 Front end processing

All speech material is quantized at 16 bits, at 8000 Hz sample rate; it is pre-emphasized with a factor of 0.97, and an energy based silence removal scheme is used. 12 MFCC + Δ vectors are obtained every 20 ms; CMV normalization is applied to reduce the influence of mismatch between training and testing acoustic conditions in a telephone environment.

To obtain the Δ feature, equation 1 is used with $h_d = 1$. The SDC *(12, P, D, k)* feature vectors are obtained by the concatenation of Δ features, as explained above. So, the dimensionality of the MFCC + Δ vector will be 24, and the dimensionality of SDC vector will be determined by *kN*, being 24, 36 and 48 for $k = 2$, 3 and 4, respectively.

The feature vectors, evaluated in the experiment were:

- MFCC + Δ vectors with *d=2 and 3*, as a baseline.
- SDC*(12, P, D, k)* vectors with a combination of :
  - *P=1, 2, 3, 4, and 5*
  - *D= 2 and 3*
  - *k=2, 3, and 4*

2.3 Ahumada database

To evaluate the MI of the SDC features and its performance in front of handset and channel mismatch in speaker verification, NIST2001 Ahumada database was used [6]. Phonologically and syllabically balanced phrases, of 8 to 10 word length each, were used as the speech sample set. Telephone sessions T1 and T3 were used to guarantee a real evaluation of SDC features in front to channel and handset mismatch. In each telephone sessions, conventional telephone line was used.

3. Mutual information

Mutual Information *(MI)* between two discrete random variables *Y* and *X* is denoted as:

$$MI(X;Y) = H(X) - H(X/Y) \tag{3}$$

where *H(X)* is the unconditional entropy of *X* and *H(X/Y)* is the mean conditional entropy of *X* given *Y* [7]. *MI* measures how much knowing one of the variables reduces the uncertainty about the other.

In speaker recognition, a speech waveform generated for a speaker *S* is used to determine his/her identity: for each frame of speech, an independent feature vector x_i is obtained using some feature extraction method, the feature vector sequence $X=\{x_1,x_2,x_3,...x_T\}$ is fed to a classifier to obtain a model λ_X , which classifies the speaker *S*. *MI* between a speaker *S* – represented by λ_X – and a feature vector *X*, obtained from his/her speech, measures how much knowing of *X* reduces the uncertainty about the identity of the speaker *S*:

$$MI(S;X) = H(S) - H(S/X) \tag{4}$$

Where the unconditional entropy of the speaker set is:

$$H(S) = -\sum_{S} p(S)\log p(S) \tag{5}$$

if $p(S) = 1/N$ then $H(S) = \log N$.

The mean conditional entropy of the speaker set due all feature vectors is:

$$H(S/X) = -\sum_{i=1}^{N} p(X_i)\sum_{S} p(S/X_i)\log p(S/X_i) \tag{6}$$

The unconditional entropy of the speaker set (equation 5) depends only on the classes *S* and does not depend on the feature vector *X,* providing an upper bound of the mutual information *MI (S; X).* The mean conditional entropy of the speaker set *S* due the feature vector *X* (equation 6) can be interpreted as the decrease in the uncertainty of the speaker's identity; with a higher dependence between *X* and *S*, the knowing of *X* reduces the uncertainty about the identity of the speaker *S*, so $H(S/X)\rightarrow 0$, and there is a higher certainty in classifying a speaker given its feature vector, and $MI(S;X) \rightarrow H(S)$.

3.1 The conditional probability density function.

To obtain a good representation of the conditional *pdf* of the speaker *S* given the correspondent feature vector *X,* the use of a well known Gaussian Mixture Modeling (GMM) [8] is proposed. For a set of *N* known speakers *S* modeled by their GMMs $\lambda_1, \lambda_2....\lambda_N$, the *a posteriori probability* of classify each known speaker λ_k, due an unknown feature vector *X* is:

$$p(\lambda_k / X) = \frac{p(X/\lambda_k)}{p(X)}\, p(\lambda_k) \tag{7}$$

This paper evaluates the *MI* between the identity of each speaker λ_k and the MFCC + Δ and SDC feature vector *X*, obtained from his respective speech, using a speaker recognition experiment. In order to obtain the maximum conditional *pdf* of each speaker λ_k, the train vectors sequence *X* used to classify the speaker, must be the same used as test vectors sequence. It insures minimum uncertainty of *S* due *X,* and minimum conditional entropy *H(S/X)*, so $MI(S;X) \rightarrow H(S)$.

3.2 Mutual Information of feature vectors.

Samples were obtained from a set of 50 speakers in session T1, concatenating the ten balanced phrases of each one of them, obtaining speech samples of about 40 sec. each. MFCC + Δ and SDC feature vectors were obtained as explained in section 2.2 and were used to obtain GMMs λ_k models for each speaker. With the same MFCC + Δ and SDC vectors, now used as test vectors, were obtained the respective conditional *pdf* (equation 7) of occurrence of each speaker.

For each one of the feature vectors mentioned in section 2.2, a matrix of 50 models λ_k versus 50 unknown vectors X_i was obtained, each one of the elements of the matrix contains the conditional *pdf* of each speaker S_k respect to each feature vector sequence X_i as shown in Table 1.

Table 1. *Matrix with the conditional pdfs.*

Spkr	***Model***	X_1	X_2	...	...	...	X_{50}
S_1	λ_1	$p(S_1/X_1)$	$p(S_1/X_2)$				$p(S_1/X_{50})$
S_2	λ_2	$p(S_2/X_1)$	$p(S_2/X_2)$				$p(S_2/X_{50})$
...	...						
...	...						
...	...						
S_{50}	λ_{50}	$p(S_{50}/X_1)$	$p(S_{50}/X_2)$				$p(S_{50}/X_{50})$

The mean conditional entropy $H(S/X)$ of the 50 speakers due the vectors X_i was obtained evaluating equation 6 for columns in table 1. Considering a normal distribution of occurrence of each feature vector, then $p(X_i) = 1/50$.

The evaluation of the *MI* (equation 4) for each feature vector mentioned in section 2.2, is shown in Table 2, noting that the parameter $k=1$ represents the MFCC + Δ vector. Taking into account that each speaker S is randomly selected with uniform probability from a set of 50 speakers, so $p(S)= 1/50 \rightarrow H(S)=\log 50 = 3.91$.

Table 2. *Evaluation of Mutual Information of SDC feature vectors for each combinations of D, P and k.*

D	**2**				**3**			
k	**1**	**2**	**3**	**4**	**1**	**2**	**3**	**4**
P=1	0.48	0.27	0.62	1.17	0.54	0.44	0.87	1.43
P=2	0.48	0.26	0.61	1.03	0.54	0.37	0.93	1.56
P=3	0.48	0.28	0.61	0.99	0.54	0.40	0.84	1.41
P=4	0.48	0.27	0.54	0.85	0.54	0.40	0.84	1.31
P=5	0.48	0.24	0.45	0.65	0.54	0.39	0.84	1.24

These previous results reflect that:

- *MI* is directly correlated with the width of *D* and the number *k* of concatenated Δ frames.
- SDC feature vectors with *k=2* present worst MI than correspondent MFCC + Δ feature vectors.
- SDC feature vectors with combinations *(12,2,P,3), (12,2,P,4), (12,3,P,3), and (12,3,P,4)* has better *MI* than correspondent MFCC + Δ feature vectors.

Last conclusion carries us to select these SDC vectors for their evaluation in a speaker verification experiment.

4. Speaker verification experiment

An evaluation of previous results was done in a speaker verification experiment using a GMM/UBM classifier [9], trained and tested with the ten balanced phrases. As models of experiment, previous models λ_k were used. To evaluate the performance of MFCC + Δ and selected SDC feature vectors in front of channel/handset mismatch, test samples were obtained now from each one of the ten phrases of the speakers, but in session T3 (about 4 sec. of speech each). All 50 speakers were used as targets for their corresponding models and as impostors for the rest of models, so were obtained 500 target and 24500 impostors. Ten balanced phrases of other 50 speakers in session T1 were used to train the UBM.

This experiment was done with MFCC + Δ feature vectors with D = 2 and 3 as baseline, and SDC feature vectors with parameter combinations *(12,2,P,3), (12,2,P,4), (12,3,P,3) and (12,3,P,4).* An experiment using SDC vectors with *k=2* was done with bad EER results, compared with MFCC + Δ vector baseline, so, it was discarded in this paper.

Evaluation of the speaker verification results was performed using DET curve [10], shown in Fig. 2, 3, 4 and 5 respectively:

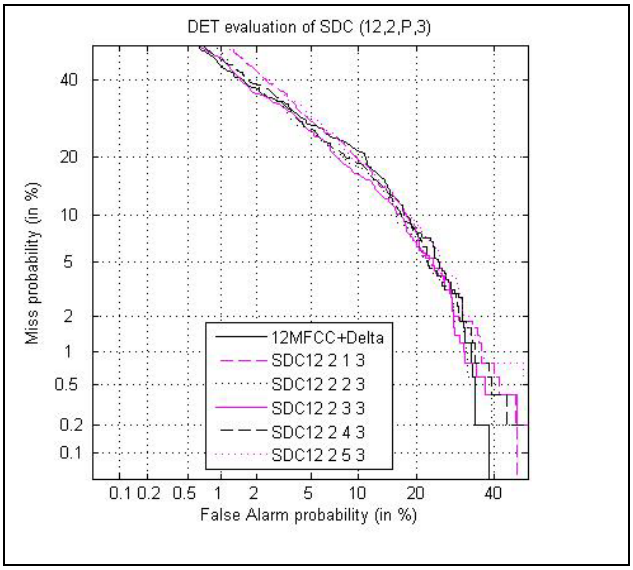

Figure 2. S*peaker verification with SDC (12,2,P,3)*

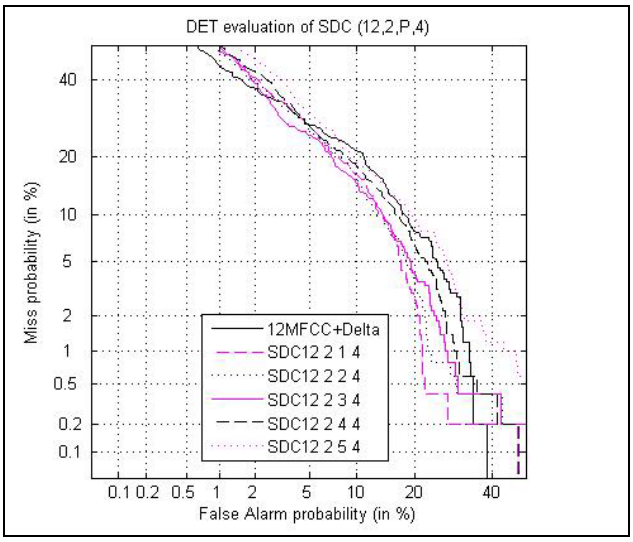

Figure 3. S*peaker verification with SDC (12,2,P,4)*

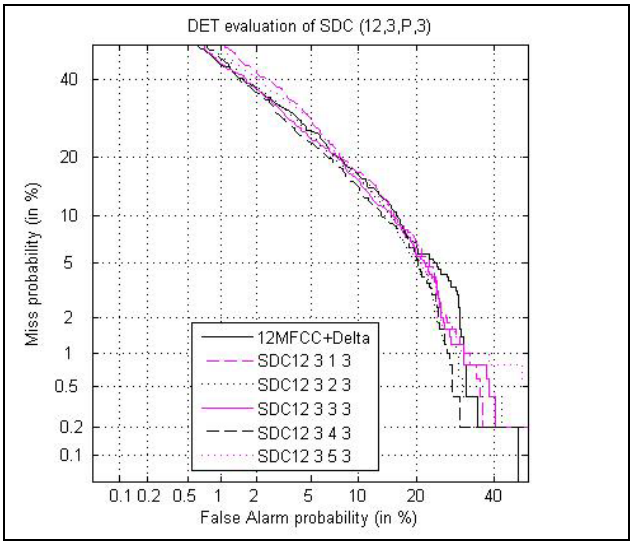

Figure 4. S*peaker verification with SDC (12,3,P,3)*

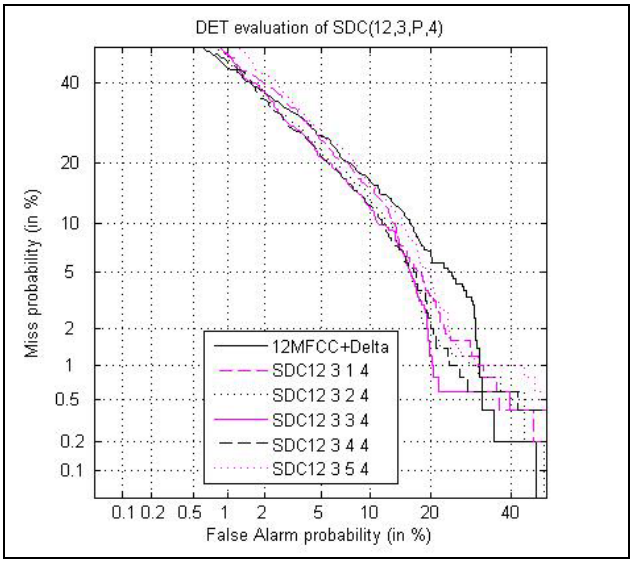

Figure 5. S*peaker verification with SDC (12,3,P,4)*

Table 3 shows the EER results; taking into account that EER for MFCC + Δ vector baseline were, for *D=2*, 14.4, and for *D=3*, 13.1, columns identified as " %" reflect the relative EER reduction of previous SDC feature vector column respect to MFCC + Δ vector baseline EER result:

Table 3. *EER results of speaker verification experiment*

D	**2**				**3**			
k	**3**	**%**	**4**	**%**	**3**	**%**	**4**	**%**
P= 1	14.	2.7	12.4	13.8	13.2	-0.1	12.3	6.1
P= 2	13.7	4.8	11.8	18.0	12.4	5.3	11.4	12.9
P= 3	13.2	8.3	12.3	14,6	12.2	6.8	10.6	19.0
P= 4	13.9	3,4	13.6	5.5	11.7	10.6	11	16.0
P= 5	14.1	2.1	14.2	1.4	12.8	2.3	12.6	3.8

DET curves and Table 3 reflect that:

- In general, SDC vectors reflect better EER than MFCC + Δ vector baseline, for any *D , P* and *k*.
- Better results were obtained with SDC vectors with *k=4,* confirming the previous MI evaluation.
- Worst results were obtained with SDC vectors with *P=5,* indicating that the excessive time shift between consecutive Δ frames reduces or disappears the necessary overlap with the time frames of Δ calculation.
- The most interesting is referred to the better EER result obtained for each combination of *D, P* and *k* (gray cells in table 3). For each *D* (time width of Δ) and *k (*number of concatenated Δ frames) there was an optimal value of *P* (time shift between consecutive Δ frames), where better EER result was obtained: if *D* increases or *k* decreases the better result was obtained with greater *P*.

5. Conclusion and future work

Speaker verification results in section 4 reflect that previous evaluation of *MI* between the identity of each speaker and his speech feature vector in section 3.2, permitted us to select the better combination of SDC parameters in front of channel/handset mismatch in speaker verification. The SDC vector with combination *(12,2,2,4*) reflects an EER relative reduction of 18 %, and SDC vector with combination *(12,3,3,4*) reflects an EER relative reduction of 19%, both respect to MFCC + Δ vector baseline.

The adequate selection of SDC parameters set *(D, P ,k)* is an important issue in any speaker recognition application, if we require less computational cost, the use of *SDC (12,2,2,4*) is better, but if we require less dimensionality, the use of *SDC(12,3,4,3*) is the better choice. The optimal relation between *D* (time width of Δ), *k (*number of concatenated Δ frames) and *P* (time shift between consecutive Δ frames) related to the better EER, must be deeply explored.

In a previous work [11] the authors evaluate the pseudo-prosodic behavior of SDC features, using a combination *SDC (12,2,2,2),* observing that SDC features reflect some correlation with prosodic features. The results presented in this work confirm that SDC features must be considered as an alternative to Δ features, without additional computational cost, in order to reduce the effects of channel/handset mismatch in speaker verification performance.

The results of this analysis and its experimental confirmation can be used to explore future applications of SDC feature of speech in speaker detection and tracking, taking into account its computational simplicity.

6. References

1. Ortega-Garcia, J., Bigun, J., Reynolds, D.,González-Rodríguez, J. Authentication gets personal with biometrics. IEEE Signal Processing Magazine. March 2004, pp 50-62.
2. Torres-Carrasquillo, P.A., Singer, E., Kohler, M.A., Greene, R.J., Reynolds, D.A., Deller J.R. Approaches to language identification using Gaussian Mixture Models and shifted delta cepstral features. Proceedings of ICSLP 2002, pp. 89-92.
3. Calvo J. R., Fernandez R. y Hernandez G. Channel/Handset mismatch evaluation in a biometric speaker verification using shifted delta cepstral features. Proceedings of CIARP 2007. LNCS 4756, pp 96-105.
4. Lu, X., Dang, J. Dimension reduction for speaker identification based on mutual information. Proceedings of Interspeech 2007, pp 2021-2024.
5. Ganchev, T., Zervas, P., Fakotakis, N., Kokkinakis, G. Benchmarking feature selection techniques on the speaker verification task. Proceedings of the Fifth International Symposium On Communication Systems, Networks And Digital Signal Processing, 2006, pp 314-318.
6. Ortega-Garcia, J., Gonzalez-Rodriguez, J., Marrero-Aguiar, V. AHUMADA: A large speech corpus in Spanish for speaker characterization and identification. Speech Communication, 2000. Vol 31, pp 255-264.
7. Gray R.M. Entropy and Information Theory. Springer-Verlag, 2007, New York.
8. Reynolds, D. A., Rose, R. C. Robust text-independent speaker identification using Gaussian Mixture speaker Models. IEEE Transactions on Speech and Audio Processing. Vol 3(1), 1995.
9. Reynolds, D. A., Quatieri T. F., Dunn R. B. Speaker verification using adapted Gaussian Mixture Models. Digital Signal Processing. Vol 10, pp 19-41, 2000.
10. Martin A., Doddington K. G., Ordowski M., Przybocki M. The DET curve in assessment of detection task performance. Proceedings of EuroSpeech 1997. Vol 4, pp 1895- 1898.
11. Calvo, J. R., Ribas, D., Fernández, R., Hernández G. Evaluation of linear relation between shifted delta cepstral features and prosodic features in speaker verification. Proceedings of CIARP 2008. LNCS 5197, 112-119.