Speaker Verification with Shifted Delta Cepstral Features: Its Pseudo-Prosodic Behavior

Dayana Ribas González, José R. Calvo de Lara

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

Abstract

This paper examines the linear relation between Shifted Delta Cepstral (SDC) features and the dynamic of prosodic features. SDC features were reported to produce superior performance to Δ features in Language Identification and speaker recognition systems. A selection of more correlated SDC features is used in speaker verification to evaluate its robustness to channel/handset mismatch. The experiment reflects superior performance of selected SDC features regarding to features in speaker verification using speech samples from NIST 2001 Ahumada database.

Index Terms: speaker verification, shifted delta cepstral, prosodic features, channel mismatch.

1. Introduction

Different studies have been done to use dynamic information contained in speech. The most popular approach consists in extracting first and second order time derivatives of instantaneous cepstral features: delta (Δ) and delta-delta ($\Delta\Delta$) features. Furui [1] used cepstral coefficients and their regression coefficients for speaker recognition, and established the effectiveness of combining temporal and dynamic features.

(Δ) and ($\Delta\Delta$) features reflect short-term speech spectral dynamics and don't capture longer term variation in speech, reflected in other 'high level' speaker dependent features, as prosodic, phonetic and linguistic [2]. But these last approach require a lot of speech samples and are also time consuming and computationally complex.

Recently the use of a longer term temporal feature called Shifted Delta Cepstral (SDC), in language recognition [3] and speaker recognition [4, 5], has improved the performance of the recognizer in front to channel and handset mismatch.

As a longer term temporal feature, SDC reflect the dynamic of the spectral features and could have a pseudo prosodic behavior. This paper explore this possibility, evaluating the linear relation between SDC features and the dynamic of two prosodic features -pitch and energy- in two different contexts -read text and free expression- and selecting a reduced set of the most correlated SDC features in a speaker recognition experiment under channel and handset mismatch conditions.

The rest of the paper is organized as follows. Section 2 describes the features. Section 3 evaluates the lineal relation between SDC features and the dynamic of pitch and energy. Section 4 describes the experiment and results. Section 5 concludes this work and gives future research direction.

2. Shifted Delta Cepstral and Prosodic Features

For efficient representation of the cepstral dynamic trajectory over some short segment of speech, Furui [1] suggested the use of an orthogonal polynomial fit of each cepstral coefficient *c(t)* trajectory over a finite length time window h_d. The 1^{st} order coefficient, or the generalized spectral slope in time, $\Delta_c(t)$ is denoted as:

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

A rectangular window (hd = 1) of reasonable length has to be used to ensure a smooth fit to the data points from one frame to the next. Δ and $\Delta\Delta$ features usually have been calculated using Eq. (1) with D between 2 to 4, depending on frame time length.

Originally proposed by Bielefeld [6], SDC features are specified by a set of 4 parameters, (N, D, P, k) where:

- N: number of coefficients in each cepstral vector.
- d: time advance and delay for the delta computation.
- P: time shift between consecutive blocks.
- k: number of blocks whose delta coefficients are concatenated to form the SDC vector

First, a N-dimension cepstral feature vector is computed in each speech frame t, then each c coefficient is differenced using spaced tD frames to obtain the Δ features, at last k different Δ features, spaced P frames apart, are stacked to form a SDC feature vector for each frame. The SDC vector at frame time t is given by the concatenation from i=0 to k-1 blocks of all the Δc(t + iP), where:

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

Eq.(2) is a generalization of eq.(1) with $h_d = 1$, including the iP time shift.

The calculation of SDC features doesn't require extra computational cost, respect to Δ features, recent experiments have shown an improvement of speaker recognition performance [4, 5] without an increase of dimensionality.

Prosodic features are considered longer term characteristics because they provide a description of the habitual attributes of the speaker. Pitch and energy have a robust performance in speaker recognition specially when dealing with noisy and mismatched channels. Besides they have speaker specific information, due to vocal folds physical differences between speakers. The unpractical aspect of prosodic features is the high amount of data needed for a successful recognition, also the procedure required to obtain them is complicated and computationally expensive [7].

Prosodic information can be used taken global statistics of the features, like mean and standard deviation of the pitch and energy. But that approach doesn't capture the temporal dynamic information of the prosodic feature. Another approach is to obtain a representation of the temporal trajectory of the pitch and energy contours. But that isn't efficient enough. Previous work had proven the utility of the derivative functions of pitch and energy in the description of their dynamic [8].

2.1. A pseudo prosodic behavior of SDC features

Dynamic Δ and $\Delta\Delta$ features, evaluated over extended speech time intervals, have been used in speaker recognition as a characteristic which contains useful additional information about speaker identity. Furui [1] recommends a time interval of 90 ms to preserve the transitional information associated with changes from one phoneme to another, Soong and Rosemberg [9] recommends a time interval from 100 to 160 ms to obtain good estimates of the trend of spectral transitions between syllables.

Alternatively SDC, as a longer term temporal feature, describes the spectral dynamic of speech. Cepstral features contain information about speech formants structure and its dynamic can reflect the movement and position of vocal and nasal articulators, if the time interval is enough longer. In each frame, SDC features reflect the temporal dynamic of the articulators in the next frames, as a pseudo-prosodic feature vector, computed without having to model the prosodic structure of the speech.

Three combinations of SDC features are proposed to obtain a good estimate of the dynamic of spectral transitions and compare the behavior of SDC and cepstral + Δ feature. The value k was fixed at 2 to ensure similar dimensionality between features. It was considered the time interval necessary and sufficient to choice the value D. Table 1 reflects used combinations of SDC features:

Table 1: *SDC features combinations.*

D	P	k	frames	time interval
2	2	2	7	147ms
2	3	2	8	168ms
3	2	2	9	189ms

This work evaluates the pseudo-prosodic behavior of SDC features through the linear relation between SDC and the dynamic of pitch and energy. Then, those SDC feature vectors more correlated, will be selected to evaluate its robustness in a telephone speaker recognition experiment.

3. Temporal relation of SDC features with prosodic features

To evaluate the lineal relation that could exists between SDC and prosodic features, this work uses the temporal correlation between a time sequence of SDC features and the dynamic of pitch and energy. Cross-correlation between two N-length sequences x and y, provides a statistical comparison of both as a function of the time-shift m and indicates the strength and direction of a linear relationship between them.

$$\Phi_{xy}[m] = \frac{1}{2N+1}\sum \frac{t=1}{N-m} x[t]y[t+m] \qquad (3)$$

If x and y are standardized, the limits of cross-correlation are $-1 \leq \Phi_{xy}[m] \leq 1$, the bounds 1 indicating maximum correlation and 0 indicating no correlation. A high negative correlation indicates a high inverse linear relation.

3.1. Cross-correlation between SDC components

Proposed combinations of SDC features (Table 1) are constituted by two blocks Δc(t) and Δc(t+2), obtained with eq.(2) evaluated at i=0, 1 with P=2,3 and D=2,3. Both blocks are highly correlated, due by the strong linear dependence between them. Cross-correlation between two consecutive blocks of any SDC vector is +1, at P distance of the lag m=0, and present maximum negative correlation in two symmetrical lags respect to P at (D+2). Figure 1 shows, in combination SDC (N, 2, 2, 2), the correlation of Δc(t), the cross correlation between both blocks of SDC, and the correlation of the mean of both blocks, all of them have the same behavior.

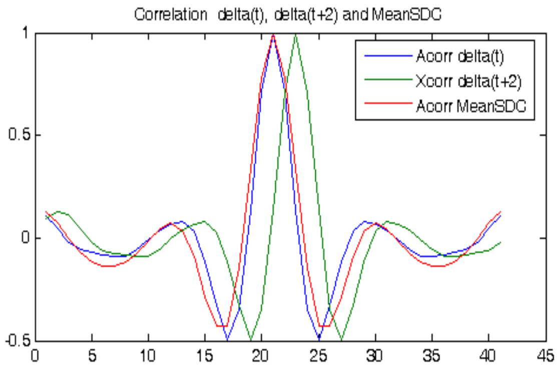

Figure 1: Cross correlation between two consecutive SDC blocks.

This property of high correlation between any two consecutive blocks is used to simplify the computation of the cross-correlation between SDC and prosodic features in this work, representing SDC feature as the mean of blocks Δc(t) and Δc(t+2).

3.2. Cross correlation between SDC and Δpitch/Δenergy

To evaluate correlation between mean SDC features and the dynamic of prosodic features, two expressions -read text and spontaneous speech- of 30 speakers of NIST2001 Ahumada database [10] were used, representing about 90 minutes of telephone speech. 12 MFCC+Δ vectors and their corresponding SDC vectors, and pitch and energy values, were synchronously obtained in each frame, to conform the time sequences. Δpitch and Δenergy were calculated using eq.(1) with D=2. Mean and variance normalization were applied as a feature standardization method.

Cross-correlation of the three proposed combinations of SDC features with Δpitch and Δenergy, presents very similar behavior respect to upper and lower peaks values and their lags positions. So, combination SDC(12,2,2,2) was selected for the experiment, as the less computationally expensive SDC feature of Table 1.

The results of cross-correlation evaluation between each one of the 12 SDC features with Δenergy and Δpitch, are showed through SDC features organized in decreasing order of correlation in Table 2. The highest correlations of SDC were obtained with respect to Δ energy. In general the correlation peaks are negative, reflecting an inverse lineal relation, it means, an increase of one time sequence implies a decrease in the other.

Although the values of the cross-correlation peaks are not very impressive, there are some SDC features more correlated than others. The most correlated values are between -0.65 and -0.35 and the rest are between -0.2 and 0.3.

Table 2: *SDC features organized in decreasing order of cross-correlation.*

order	**1**	**2**	**3**	**4**
Δ E	sdc_4	sdc_5	sdc_3	sdc_2
xcorr	-0.65	-0.56	-0.55	-0.45
Δ P	sdc_4	sdc_6	sdc_5	sdc_3
xcorr	-0.67	-0.50	-0.48	-0.37
order	**5**	**6**	**7**	**8**
Δ E	sdc_6	sdc_9	sdc_{11}	sdc_8
xcorr	-0.37	-0.35	-0.27	-0.25
Δ P	sdc_9	sdc_7	sdc_1	sdc_2
xcorr	-0.35	-0.35	0.35	-0.30
order	**9**	**10**	**11**	**12**
Δ E	sdc_{10}	sdc_{12}	sdc_7	sdc_1
xcorr	-0.25	-0.25	-0.22	0.12
Δ P	sdc_{11}	sdc_{12}	sdc_{10}	sdc_8
xcorr	-0.3	-0.27	-0.2	-0.18

Then, two vectors of six SDC features were used in speaker verification experiment, appended to MFCC vector, the first vector, more correlated with Δenergy, composed by sdc_2, sdc_3, sdc_4, sdc_5, sdc_6 and sdc_9 and a second vector, more correlated with Δpitch, composed by sdc_3, sdc_4, sdc_5, sdc_6, sdc_7 and sdc_9. Both resultant vectors have the same dimensionality as MFCC+Δ vector.

3.3. Experiments and Results

NIST 2001 Ahumada [10] is a speech database of 103 Spanish male speakers, acquired under controlled conditions for speaker characterization and identification. A speaker verification experiment is performed using ten phonologically and syllabically balanced phrases in telephone sessions.

Training samples set is obtained under good handset/channel characteristics, concatenating the ten balanced phrases (about 40 sec. of speech) of each one of 50 client speakers. Testing samples sets are obtained with each one of the phrases of the same speakers in another session (about 5 sec. of speech each), made using 9 randomly selected standard handsets and each speaker uses one of them.

For each handset, three characteristics were reported: (a) microphone sensibility, (b) microphone band pass frequency response, and (c) signal to noise ratio in its associated channel. Test was performed with samples of those 50 clients who speak under the worst mismatch condition, in order to evaluate the robustness in front to channel mismatch due to:

- low microphone handset sensibility($< 1mV/P$)
- low microphone band pass frequency response($< 20dB$)
- low signal to noise ratio in the channel($< 30dB$)

Each frame of speech is represented by a 12-dimensional MFCC features vector. Cepstral Mean and Variance Normalization feature normalization method is applied to MFCC features. The Δ cepstral vector is obtained from each cepstral feature using Eq.(1) with D=2. The SDC(12,2,2,2) vector is obtained concatenating one additional Δ cepstral vector separated P=2 spaces, to original Δ cepstral vector. This work evaluates the behavior of the two selected SDC feature vectors (epig. 3.2) appended to MFCC vector, respect to MFCC + Δ vector. So, three different sets of features with the same dimensionality, are used in the experiment:

1. 12 MFCC+12 Δ, dimension 24 (baseline): M-D
2. 12 MFCC+6 SDC more correlated with Δenergy: M-SDC-E
3. 12 MFCC+6 SDC more correlated with Δpitch: M-SDC-P

The experiment performance is evaluated using a 64 mixtures GMM/UBM classifier [11], trained and tested with the ten balanced phrases of 50 client speakers of the database. The ten balanced phrases of other subset of 50 non client speakers are used to train the 256 mixtures UBM.

Experiment results are reflected in detection error tradeoff (DET) plots:

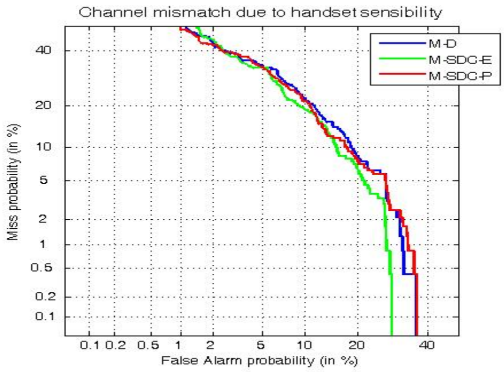

Figure 2: Speaker verification under low microphone handset sensibility ($< 1mV/P$).

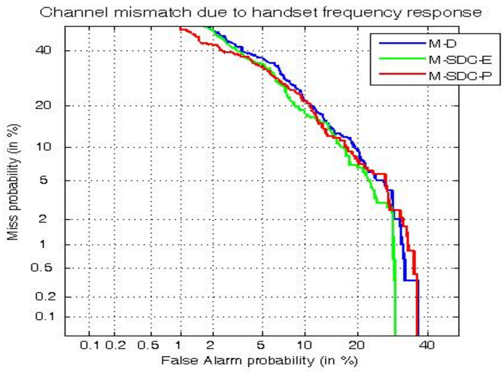

Figure 3: Speaker verification under low microphone band pass frequency response ($< 20dB$).

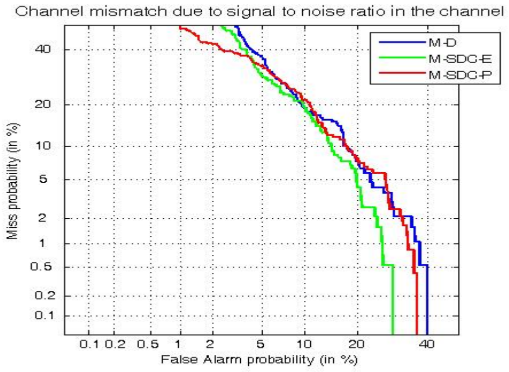

Figure 4: Speaker verification under low signal to noise ratio in the channel ($< 30dB$).

The values of EER and DCF of the experiments are showed in Table 3. Then Table 4 reflects the relative reduction in percent of EER, for both sets of SDC features.

Table 3: *EER and DCF results of baseline and two set of selected SDC features.*

Set of features		Low handset sensibility	Low handset freq. response	Low s/n ratio
M-D	EER	14.7	14.2	15.3
	DCF	0.06	0.065	0.068
M-SDC	EER	13.7	13.9	12.6
-E	DCF	0.061	0.066	0.069
M-SDC	EER	13.2	13.4	13.4
-P	DCF	0.068	0.062	0.058

Table 4: *Reduction in percent of EER for both sets of selected SDC features respect to baseline.*

Mismatch condition	SDC correlated with Δ energy	SDC correlated with Δ pitch
low handset sensibility	6.8	8.8
low handset freq. response	2.1	5.6
low s/n ratio in channel	17.6	13.7

4. Conclusions and Future Work

This work reflects the results obtained in the evaluation of a prosodic-related vector of SDC features, in speaker verification using speech samples from mismatch telephone sessions of NIST2001 Ahumada database.

Results in DET plots of speaker verification experiments reflect:

- a superior performance respect to MFCC + Δ features of both prosodic-related SDC features (see Table 3).
- a better performance respect to MFCC + Δ features, of SDC features more correlated with Δ energy (see figures 2, 3 and 4). This result is consistent with the highest correlation of SDC features with Δ energy (epig. 3.2)
- a superior robustness of both prosodic-related SDC features, mainly under low s/n in the channel, consistent with robustness of prosodic features (see Table 4).

As SDC features reflect correlation with prosodic features, without additional cost respect to Δ features, they must be considered as an alternative to Δ features, in order to reduce the effects of channel/handset mismatch in speaker verification performance.

Future work will be in the direction of evaluate another relations between SDC features and the dynamics and statistic of prosodic features.

5. References

[1] S. Furui, 'Cepstral analysis for automatic speaker verification', IEEE Transactions on ASSP, 29(2):254-272, 1981.

[2] D. Reynolds and W. Andrews and J. Campbell and J. Navratil and B. Peskin and A. Adami and Q. Jin and D. Klu-sacek and J. Abramson and R. Mihaescu and J. Godfrey and D. Jones and B. Xiang.Stone and H.S., "The SuperSID Project: Exploiting High-level Information for High-accuracy Speaker Recognition", Proceedings of the IEEE ICASSP 2003, vol. 4:784-787.

[3] P. A. Torres-Carrasquillo and E. Singer and M.A. Kohlerand R. J. Greene and D.A. Reynolds and J.R. Deller Jr. "Approaches to language identification using Gaussian Mixture Models and shifted delta cepstral features." Proceedings of ICSLP 2002, pp. 89-92

[4] T. Kinnunen, C.W. E. Koh, L. Wang, H. Li, E. S. Chang. Temporal discrete cosine trans-form: Towards longer term temporal features for speaker verification, Proceedings of ICSLP 2006.

[5] J. Calvo and R. Fernndez and G. Hernndez. "Channel/Handset Mismatch Evaluation in a Biometric Speaker Verification using Shifted Delta Cepstral Features." Proceedings of CIARP 2007, LNCS 4756, pp.96-105.

[6] Bielefeld. B. "Language identification using shifted delta cepstrum." Proceedings Four-teenth Annual Speech Research Symposium 1994.

[7] L. Mary and B. Yegnanarayana. "Prosodic features for Speaker Verification" Proceedings of Interspeech 2006.

[8] A. Adami and R. Mihaescu and D. Reynolds and J. Godfrey. "Modeling prosodic dynamics for Speaker Recognition." Proceedings of ICASSP 2003.

[9] F. Soong and A. Rosenberg. "On the use of instantaneous and transitional spectral information in speaker recognition." IEEE Trans on Audio Speech and Signal Proc. 36(6):871-879, 1988.

[10] J. Ortega-Garcia and J. Gonzalez-Rodriguez and V. Marrero-Aguiar. "AHUMADA: A large speech corpus in Spanish for speaker characterization and identification." Speech Comm. 31:255-264, 2000.

[11] D. Reynolds and T. Quatieri and R. Dunn. "Speaker Verification Using Adapted Gaussian Mixture Models." Digital Signal Proc. 10:19-41, 2000.