EFFECTIVE USE OF DCTS FOR CONTEXTUALIZING FEATURES FOR SPEAKER RECOGNITION

Mitchell McLaren, Nicolas Scheffer, Luciana Ferrer, Yun Lei

Speech Technology and Research Laboratory, SRI International, California, USA

{mitch,scheffer,lferrer,yunlei}@speech.sri.com

ABSTRACT

This article proposes a new approach for contextualizing features for speaker recognition through the discrete cosine transform (DCT). Specifically, we apply a 2D-DCT transform on the Mel filterbank outputs to replace the common Mel frequency cepstral coefficients (MFCCs) appended by deltas and double deltas. A thorough comparison of algorithms for delta computation and DCT-based contextualization for speaker recognition is provided and the effect of varying the size of analysis window in each case is considered. Selection of 2D-DCT coefficients using a zig-zag approach permits definition of an arbitrary feature dimension using the most energized coefficients. We show that 60 coefficients computed using our approach outperforms the standard MFCCs appended with double deltas by up to 25% relative on the NIST 2012 speaker recognition evaluation (SRE) corpus in both Cprimary and equal error rate (EER) while additional coefficients increase system robustness to noise.

Index Terms— Contextualization, Deltas, 2D-DCT, Filterbank Energies, Speaker Recognition

1. INTRODUCTION

Cepstral-based speech features are typically based on short-time analysis of speech. Consequently, they do not contain sufficient information pertinent to the manner in which the speaker uttered the phone or word to which the extracted feature belongs. For this reason, these features are typically appended with deltas and double deltas (denoted as $\triangle\triangle$) to provide context and improve performance in the fields of speech processing [1, 2]. In automatic speech recognition (ASR) and language identification (LID), contextualization through the discrete cosine transform (DCT) has proven successful [3, 4, 5, 2, 6].

The DCT breaks down an input signal into a collection of cosine functions with corresponding coefficients (or weights) that can be used to maximally reconstruct the signal. Deltas, on the other hand, make no attempt to retain information for reconstruction, and instead observe the change in values over an analysis window. Consequently, a large amount of content that is available in the analysis window is compressed through deltas; DCTs attempt to retain this information. The additional detail available from feature contextualization through DCTs therefore warrants thorough investigation in the context of speaker identification (SID).

In this paper, we investigate the contextualization of short-term speech features using the two-dimensional DCT (2D-DCT) in the context of speaker identification. We show that the selection of DCT coefficients as used in speech and language recognition performs poorly compared to several common algorithms for delta calculation. We propose a modified zig-zag selection technique, borrowed from the field of image recognition [7], in which coefficients are selected from a rectangular DCT matrix to a spectrotemporal profile from the speech. Through a series of experiments using a subset of the PRISM dataset [8], we illustrate the effect of analysis window and feature dimension of both MFCC+$\triangle\triangle$ and 2D-DCT coefficients in the context of both clean and noisy speech. Using the zig-zag selection regime, we present experiments illustrating that additional DCT coefficients in place of raw MFCCs provides improved SID performance. The tuned features are then evaluated on the recent NIST SRE'12 corpus using a state-of-the-art i-vector configuration to highlight the benefits of the proposed feature.

2. CONTEXTUALIZATION OF FILTERBANK ENERGIES

This section describes methods of contextualizing MFCCs by appending deltas or Mel filterbank outputs using 2D-DCT coefficients. In each case, we start with Mel filterbank energies extracted every 10ms from 25ms of audio using 24 Mel filters spanning a 200-3300 Hz bandwidth. While the energies are used directly for 2D-DCT contextualization, delta-based approaches take MFCCs as input which are computed as the DCT of the log filterbank energies and retaining the first C coefficients, including c0 (typically $C = 20$).

2.1. Delta Contextualization

MFCCs are the most common feature used for SID [1]. For the last decade, deltas and double deltas ($\triangle\triangle$) have been appended to these features to provide approximately 20% relative improvement in SID error rates over the use of MFCCs alone (see Section 4.1). In this study we compare three different techniques for delta computation:

- Two-point difference (TPD): The most simple approach to deltas is to use the difference between two points equidistant from the point of interest that is central to the window of analysis. For the $d'\text{th}$ dimension of features X at time t, the delta value $\triangle_d(t) = X_d(t+l) - X_d(t-l)$, where the analysis window length $N = 2l + 1$. Note that in this approach, values within the window itself are not utilized.
- Least-squares fit (LSF): Given a window of N samples around $X_d(t)$, deltas are calculated as the least square lin-

This material is based on work supported by the Defense Advanced Research Projects Agency (DARPA) under Contract D10PC20024. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the view of DARPA or its contracting agent, the U.S. Department of the Interior, National Business Center, Acquisition & Property Management Division, Southwest Branch. The views expressed are those of the authors and do not reflect the official policy or position of the Department of Defense or the U.S. Government. "A" (Approved for Public Release, Distribution Unlimited)

978-1-4799-2893-4/14/$31.00 ©2014 IEEE

ear fit to the samples. In practice this is done by convolving the samples with a linear impulse response from $-l$ to $+l$.

- Filter (FILT): Deltas are calculated as the output of the signal processed by a filter defined by the window [-0.25, -0.5, -0.25, 0, 0.25, 0.5, 0.25]. Here, the length of the analysis window $N = 7$, and forms the approach that has used in SRI's SID and LID systems for the past decade. Increasing the length of the analysis window involves inserting zeros at the center point of the window.

Note that each of these techniques can be expressed as different transfer functions of a time-domain filter. Once computing deltas of an input signal, double deltas can be computed through reapplication of the same delta technique to the signal composed by the deltas. A comparison of these techniques and their sensitivity to the size of analysis window is given in Section 4.1.

2.2. 2D-DCT Contextualization

The use of DCTs for contextualization (that is, across time instead of across frequencies as done for MFCCs) is commonplace in the field of speech recognition [3, 5, 6] and has shown to be effective for LID [4]. Recent advances in noise-robust speech activity detection (SAD) and LID have highlighted the benefits of using modulation features computed via the DCT over the use of deltas and shifted delta cepstrum [9, 2]. These studies extract coefficients from a DCT matrix computed over a window of frames from short-term features such as MFCCs. We note that the 2D-DCT matrix of log Mel filterbank energies is equivalent to taking the DCT of MFCCs over a window of frames if all cepstral coefficients are selected from the filter bank. The challenge is then to select the most useful elements from the DCT matrix of coefficients for the purpose of SID.

In our recent work on noise-robust LID [2], a DCT matrix was computed from 7-dimensional MFCCs over $N = 41$ frames, thus encompassing a larger context than required for SID. The first row corresponding to the average of MFCCs was discarded and rows 1–9 were appended to the MFCC feature that was central to the analysis window. We replicate this DCT selection strategy for SID, however, due to a greater number of coefficients in SID (20 as opposed to 7 for LID), we restrict coefficients to rows 1 and 2 so as to produce 40 dimensions to be appended to the MFCCs, thus being comparable in dimensionality to deltas and double deltas. This rectangular parsing strategy is termed 'MFCC+DCTrec' for this study.

The sub-rectangle of coefficients selected via DCTrec places more emphasis on the frequency axis as opposed to time, where coefficients are less robustly estimated as the corresponding frequency increases. Desired is a selection strategy that captures a robust spectrotemporal profile from speech. Selecting the more robustly estimated DCT coefficients from the low-frequency bands is one means of accomplishing this. To this end, we borrow the technique of zig-zag selection from the related field of image processing [7], where it facilitates entropy encoding. In face recognition, the use of 2D-DCTs and the zig-zag parsing strategy is commonplace for GMM- and HMM-based recognition systems where coefficients are selected from square blocks of pixels values [10, 11]. We adapted this parsing strategy to accommodate a *rectangular* DCT matrix, which is dependent on the both the number of filter banks and analysis frames computed from speech frames. The devised selection strategy first constructs an outer product matrix of vectors $F = \left[\frac{f}{f}, \frac{f-1}{f}, \ldots, \frac{2}{f}, \frac{1}{f}\right]$ and $W = \left[\frac{w}{w}, \frac{w-1}{w}, \ldots, \frac{2}{w}, \frac{1}{w}\right]$ where f is the number of filterbanks and $w = N - 1$ is the size of the analysis window after removing the first row of the 2D-DCT which represents the mean over the analysis window (the inclusion of which was found to reduce SID performance). The coordinates of the C highest values in the outer product matrix are then used to select the 2D-DCT coefficients for use in SID. Figure 1 provides an example of the coefficients selected when $C = 60$. While not explored in this work, it is worth noting that the ratio of the number of filterbanks and the number of frames in the window will alter the parsing and this warrants future exploration. DCTs selected using this technique will be referred to as DCT$_{zz}$. Section 4.2 compares the use of appending DCT$_{zz}$ coefficients to MFCC features versus their use as a standalone feature.

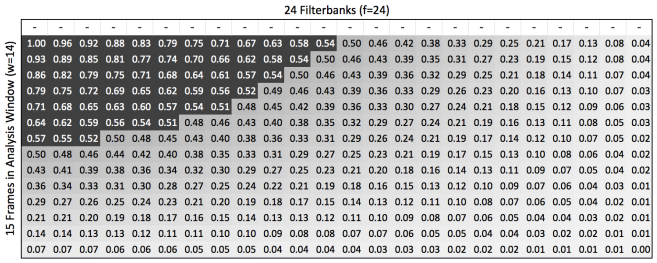


Fig. 1. Dark blocks indicate coordinates of the 60 2D-DCT coefficients selected based on the devised zig-zag parsing strategy after skipping the first row.

3. EXPERIMENTAL PROTOCOL AND SYSTEM CONFIGURATION

Two main protocols and system configuration combinations were used in this study. Both use simple GMM-based speech activity detection as defined in [12] and an i-vector/probabilistic linear discriminant analysis (PLDA) framework [13, 14].

Simplified PRISM: For the purpose of tuning, a small-scale, gender-independent system [15] was used. This system based on the PRISM protocols [8]. A diagonal covariance universal background model (UBM) with 512 Gaussian components was trained using a subset of 3220 samples from the original PRISM training list while the 400-dimensional i-vector subspace was trained using 6440 samples. Evaluation was performed on both non-degraded audio (sre10) lists and noisy (noi) lists to evaluate any trade-off between clean speech performance and noise-robustness during feature tuning. For rapid evaluation turnaround, only a subset of clean trials were used consisting of 14080 target and 688125 impostor trials from 2483 single-segment models and 3824 test segments. Performance is reported in terms of equal error rate (EER) and the minimum decision cost function (DCF) defined in the SRE'08 evaluation protocol [16].

Full SRE'12 System: A gender-dependent system was trained based on the protocols used in the development of our SRE'12 submission [12] in order to evaluate the tuned features. The number UBM components increased to 2048 and was trained using a subset of 8000 clean speech samples; the i-vector subspace was trained using 51224 samples from which 600D i-vectors were extracted. PLDA and 300-dimensional LDA for i-vector reduction was trained using using a similar extended dataset of 62277 samples (26k of which were re-noised). Evaluation was performed on female trials of the five conditions defined by NIST based on the extended protocol with performance reported in terms of Cprimary [17] and EER.

4. RESULTS

The techniques for delta calculation detailed in Section 2.1 are first compared across varying analysis window sizes. The use of 2D-DCT

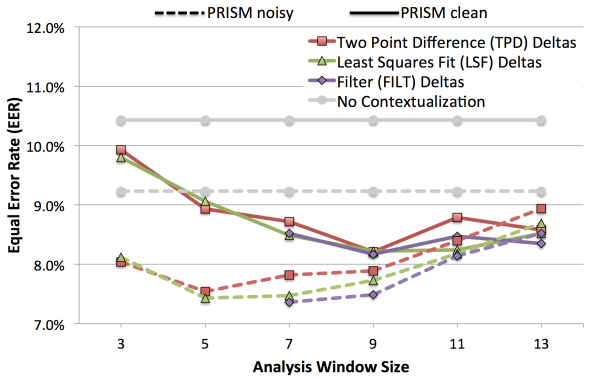


Fig. 2. Analyzing three delta calculation algorithms on the subset of clean and noisy PRISM data (solid and dashed lines respectively) with various analysis window sizes using 20D MFCCs+△△.

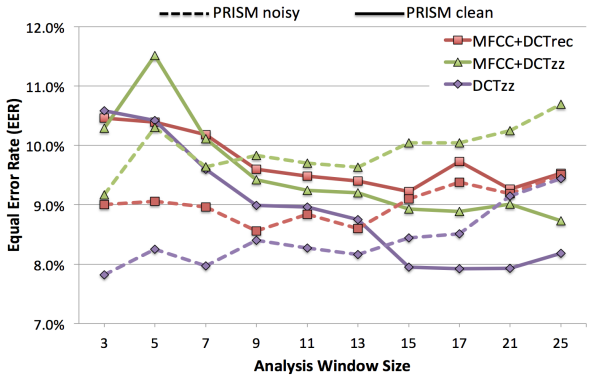


Fig. 3. Comparing DCT contextualization with and without MFCCs using various analysis window sizes on the subset of clean and noisy PRISM data with 60-dimensional features.

coefficients to replace delta-based contextualization is then evaluated along with their associated benefits over MFCCs.

4.1. Evaluating Delta Contextualization

This section compares different algorithms for delta calculation across different sizes of analysis window using MFCCs appended with △△. The motivation for this comparison is documentation of the parameter tuning of this process in the context of the modern i-vector/PLDA speaker recognition framework, and for a fair comparison to the proposed use of 2D-DCT coefficients. Figure 2 plots the EER obtained on the PRISM protocol from these experiments.

The baseline system without contextualization (i.e., 20D raw MFCCs) offers an EER of 10.4% and 9.2% in the clean and noisy conditions respectively[1]. All results in Figure 2 surpass these to varying degrees. The three methods of delta calculation are comparable in clean conditions, similarly varying with the size of analysis window. At windows beyond 9 frames, TPD tended to reduce performance compared to LSF and FILT. More variation between techniques were observed in the noisy speech trials, in which FILT deltas offered the best performance with an analysis window of 7 frames. TPD was the poorest choice for noise robustness while LSF was comparable to FILT at larger window sizes. Use of FILT with an analysis window of 9 frames offered the greatest average relative improvement of 20% over the baseline MFCC system, when averaged for clean and noisy conditions, and will be used for delta-based contextualization for the remainder of this study.

4.2. Evaluating 2D-DCT Contextualization

Experiments in this section evaluate the use of several contextualization techniques based on the 2D-DCT matrix of log Mel filterbank energies in place of △△ computation. The three techniques evaluated are 40-dimensional MFCC+DCT$_{rec}$ and MFCC+DCT$_{zz}$ in which coefficients are appended to 20D MFCC features, and 60-dimensional DCT$_{zz}$ coefficients extracted directly from the log Mel filterbank outputs without raw MFCCs. Results across various analysis window sizes are given in Figure 3.

Figure 3 illustrates that the proposed DCT$_{zz}$ features provided the best performance in both clean and noisy conditions. The improvement of DCT$_{zz}$ coefficients over coefficients appended to MFCC was more evident in the case of noisy speech. Interestingly, appending 2D-DCT coefficients to MFCCs provided little if any improvement over the baseline MFCCs in noisy conditions. The optimal window size for clean speech was between 15 and 21; noisy trials showed less sensitivity to a window size of less than 21. The best average improvement of 16% over the baseline system was found using a window size of 15 for the 2D-DCT matrix. This is comparable to the optimal context of deltas when considering that the optimal delta window of 9 is equivalent to 17 when taking into account the total number of frames needed to append double deltas.

Figures 2 and 3 show that both deltas and DCT contextualization offer the same trend of preference for larger analysis windows for clean speech and smaller windows for noisy speech. Performance on noisy speech from delta-based contextualization in Figure 2 is considerably better than the DCT-based techniques. Results in these figures are based on a fixed dimensionality of 60. The following section explores whether this is optimal for all techniques.

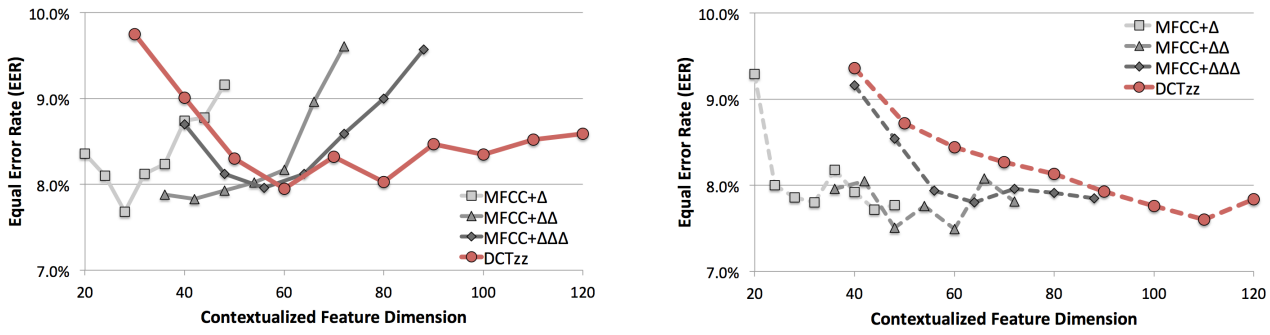
4.3. Coefficient Counts

When using delta techniques to contextualize MFCCs, one first has to select both the number of coefficients and the number of delta processes. In the proposed DCT$_{zz}$ features, increasing the number of extracted coefficients relates to more information relevant to the reconstruction of the log Mel filterbank outputs from the 2D-DCT. Thus an arbitrary selection of coefficients is possible.

In this section, we vary the feature dimension from DCT$_{zz}$ (window size 15) to observe effects on SID performance on clean and noisy data. For a thorough comparison, raw MFCCs along with different delta contextualizations are evaluated including triple deltas (△, △△, △△△). Figure 4 details the results. On clean speech in Figure 4(a), we note a general preference for lower dimensionality, with MFCC+△ offering the best performance with as few as 28 dimensions. The proposed DCTzz approach was comparable to triple deltas with the best operating dimensionality at 60 dimensions. In contrast to clean speech, noisy trials found that additional contextualization through double deltas was beneficial. Furthermore, Figure 4(b) illustrates a trend toward increased dimensionality. This trend was particularly the case for the DCTzz with as many as 110 dimensions providing best performance for this feature compared to 48 for MFCC+△△. Both MFCC+△△ and DCTzz illustrated improved noise-robustness with a feature dimension beyond the optimal for clean speech.

The use of 60 coefficients provided the best clean speech performance from DCTzz coefficients and 42 dimensions from

[1]Noisy trials are based on re-noised (and some clean) microphone segments attributing to the apparent performance gain over clean segments in which miscalibration between tel and int segments exists.



(a) PRISM Clean (b) PRISM Noisy

Fig. 4. Performance of delta and 2D-DCT contextualizations as the feature dimension varies on subset of clean and noisy PRISM data.

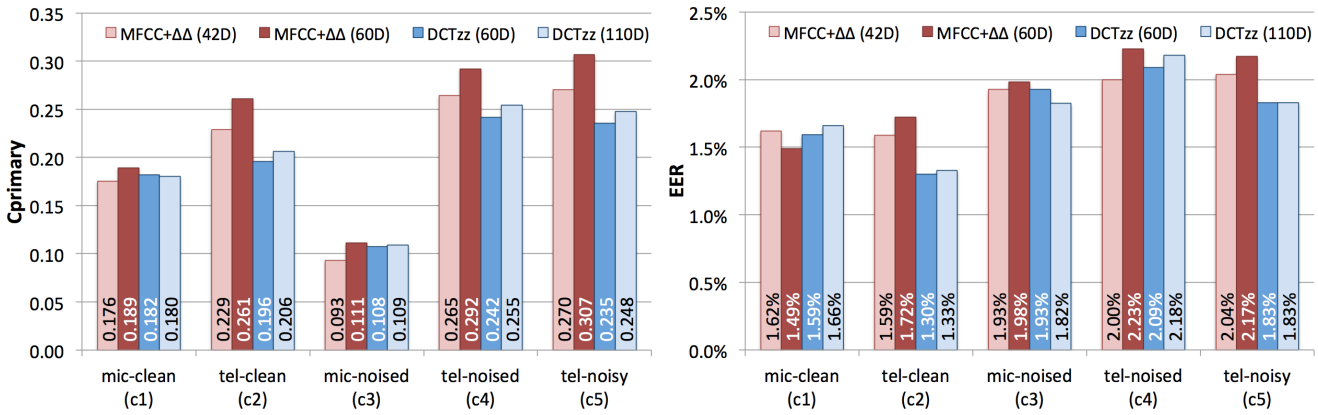


Fig. 5. Performance on conditions c1–5 of SRE'12 female trials using MFCC+△△, and 2D-DCT coefficients with different dimensions.

MFCC+△△. In contrast, noisy speech trials found benefit from the use of more coefficients: 60 dimensions for MFCC+△△ and 110 for DCTzz. The following section aims to determine the extent to which this sensitivity to dimensions exists on a large system.

4.4. NIST SRE'12 Evaluation

This section aims to determine whether the tuned analysis window and coefficient counts generalized[2] to the evaluation of the SRE'12 dataset on a large scale system as defined in Section 3. Figure 5 details results using the best delta and DCT configurations (MFCC+△△ and DCTzz, respectively) with feature dimensionality tuned on both clean and noisy speech.

Comparing 14D and 20D MFCC+△△ performance indicates that the additional raw features provided no additional robustness to noise, as anticipated based on findings in Section 4.3. This does, however, reflect the findings of [18]. The same observation can be made for the proposed DCTzz coefficients for which 110D coefficients did not improve on 60D for the noisy conditions, however the DCTzz system appears to be less sensitive to the number of coefficients than MFCC+△△. In contrast to the small development corpus used in previous sections, 60 DCTzz coefficients provided a significant 25% relative improvement in Cprimary and EER for clean telephone speech over the same dimensionality MFCC+△△. Similarly, Cprimary improvements in renoised (c4) and noisy (c5) telephone speech were 15% and 21%, respectively. Clean and noisy microphone trials were, however, comparable between MFCC+△△ and DCTzz with 60-dimensional features. When comparing both features tuned on clean speech, 42-dimensional MFCC+△△ offered some improvement over DCTzz of 60 dimensions for microphone trials in terms of Cprimary. The proposed approach, however, maintained Cprimary improvements of 9-14% relative to MFCC+△△ on telephone speech.

Two unexpected trends were observed in the above results: reduced MFCC+△△ dimensionality improved almost every condition of SRE'12 over the state-of-the-art 60-dimensional configuration and additional coefficients did not improve noise-robustness as anticipated based on feature tuning results. One hypothesis for these findings is the additional use of re-noised data in the PLDA and LDA models of the large system which has been shown to provide improved noise-robustness from 60-dimensional features [12]. Future research will determine the strength of this hypothesis.

5. CONCLUSIONS

We proposed the use of 2D-DCT coefficients from log Mel filterbank outputs to replace the widespread use of MFCCs with appended deltas and double deltas. Based on the success of DCTs for contextualization in speech processing and a coefficient selection scheme to capture information attaining to spectrotemporal profiles, the proposed 2D-DCT approach provided considerable improvements over the state-of-the-art MFCC+△△ system on clean and noisy telephone speech with 60 dimensions in the context of the recent NIST SRE'12 dataset. The proposed features exhibited less sensitivity to the choice of dimensions compared to the MFCC counterpart. Future work will improve selection of coefficients with particular regard to the preference of frequency versus time from the 2D-DCT matrix.

[2]A term to be used lightly given the bulk of SRE'12 data exists in the PRISM set, however the noises of SRE'12 when not observed during tuning.

6. REFERENCES

[1] Frédéric Bimbot, Jean-François Bonastre, Corinne Fredouille, Guillaume Gravier, Ivan Magrin-Chagnolleau, Sylvain Meignier, Teva Merlin, Javier Ortega-García, Dijana Petrovska-Delacrétaz, and Douglas A Reynolds, "A tutorial on text-independent speaker verification," *EURASIP journal on applied signal processing*, vol. 2004, pp. 430–451, 2004.

[2] Aaron Lawson, Mitchell McLaren, Yun Lei, Vikramjit Mitra, Nicolas Scheffer, Luciana Ferrer, and Martin Graciarena, "Improving language identification robustness to highly channel-degraded speech through multiple system fusion," in *Proc. Interspeech*, 2013.

[3] Climent Nadeu, Dušan Macho, and Javier Hernando, "Time and frequency filtering of filter-bank energies for robust HMM speech recognition," *Speech Communication*, vol. 34, no. 1, pp. 93–114, 2001.

[4] Fabio Castaldo, Emanuele Dalmasso, Pietro Laface, Daniele Colibro, and Claudio Vair, "Language identification using acoustic models and speaker compensated cepstral-time matrices," in *Proc. IEEE ICASSP*, 2007, pp. 1013–1016.

[5] Qifeng Zhu and Abeer Alwan, "An efficient and scalable 2D DCT-based feature coding scheme for remote speech recognition," in *Proc. IEEE ICASSP*, 2001, pp. 113–116.

[6] Frantisek Grezl and Petr Fousek, "Optimizing bottle-neck features for LVCSR," in *Proc. IEEE ICASSP*, 2008, pp. 4729–4732.

[7] WB Pennebaker and JL Mitchell, *JPEG still image data compression standard. 1993*, Van Nostrand Reinhold, New York, 2006.

[8] Luciana Ferrer, Harry Bratt, Lukas Burget, Honza Cernocky, Ondrej Glembek, Martin Graciarena, Aaron Lawson, Yun Lei, Pavel Matejka, Olda Plchot, et al., "Promoting robustness for speaker modeling in the community: The PRISM evaluation set," in *Proc. NIST 2011 Workshop*, 2011.

[9] M. Graciarena, A. Alwan, D. Ellis, H. Franco, L. Ferrer, J. H.L. Hansen, A. Janin, B.S. Lee, Y. Lei, V. Mitra, N. Morgan, O. Sadjadi, T. Tsai, N. Scheffer, L. Tan, and B. Williams, "All for one: Feature combination for highly channel-degraded speech activity detection," in *Proc. Interspeech*, 2013.

[10] Ara V Nefian and Monson H Hayes III, "Hidden Markov models for face recognition," in *Proc. IEEE ICASSP*, 1998, vol. 5, pp. 2721–2724.

[11] Roy Wallace, Mitchell McLaren, Christopher McCool, and Sebastien Marcel, "Cross-pollination of normalization techniques from speaker to face authentication using Gaussian mixture models," *IEEE Trans. on Information Forensics and Security*, vol. 7, no. 2, pp. 553–562, 2012.

[12] L. Ferrer, M. McLaren, N. Scheffer, Y. Lei, M. Graciarena, and V. Mitra, "A noise-robust system for NIST 2012 speaker recognition evaluation," in *Proc. Interpseech*, 2013.

[13] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Trans. on Speech and Audio Processing*, vol. 19, pp. 788–798, 2011.

[14] S.J.D. Prince and J.H. Elder, "Probabilistic linear discriminant analysis for inferences about identity," in *Proc. IEEE International Conference on Computer Vision*. IEEE, 2007, pp. 1–8.

[15] Mohammed Senoussaoui, Patrick Kenny, Niko Brummer, Edward De Villiers, and Pierre Dumouchel, "Mixture of PLDA models in i-vector space for gender independent speaker recognition," in *Proc. Int. Conf. on Speech Communication and Technology*, 2011.

[16] *The NIST Year 2008 Speaker Recognition Evaluation Plan*, 2008, `http://www.itl.nist.gov/iad/mig/tests/sre/2008/sre08_evalplan_release4.pdf`.

[17] *The NIST Year 2012 Speaker Recognition Evaluation Plan*, 2012, `http://www.nist.gov/itl/iad/mig/upload/NIST_SRE12_evalplan-v17-r1.pdf`.

[18] D. Colibro, C. Vair, K. Farrell, G. Karvitsky, N. Krause, S. Cumani, and P. Laface, "Nuance - Politecnico di Torino's 2012 NIST speaker recognition evaluation system," in *Proc. Interspeech*, 2013.