Odyssey 2012
The Speaker and Language Recognition Workshop
25-28 June 2012, Singapore

Exemplar-based Sparse Representation and Sparse Discrimination for Noise Robust Speaker Identification

Rahim Saeidi†, Antti Hurmalainen⋆, Tuomas Virtanen⋆, David A. van Leeuwen†

†Centre for Language and Speech Technology, Radboud University Nijmegen, The Netherlands
⋆Department of Signal Processing, Tampere University of Technology, Tampere, Finland
{r.saeidi, d.vanleeuwen}@let.ru.nl {antti.hurmalainen, tuomas.virtanen}@tut.fi

Abstract

Probabilistic modeling is the most successful approach widely used in speaker recognition either for modeling the speakers in GMM-UBM structure or by serving as a prior in secondary-level feature extraction to form i-vectors. In this paper, we introduce exemplar-based sparse representation and sparse discrimination for closed-set speaker identification in a noisy living room from very short speech segments each of 2 seconds length on average. Large spectro-temporal contexts in mel-frequency band energy domain are used to build dictionary of all speakers and decomposing the observed noisy speech, the sparse activations are extracted as features for modeling stage. Sparse discriminant analysis is employed to learn sparse discriminative directions for classification stage. Experiments on the recently developed *computational hearing in multi source environments* (CHiME) corpus demonstrate excellent performance of the proposed approach specially in low-SNR. The speaker identification results are also reported for baseline text-independent GMM-UBM and text-dependent HMM.

1. Introduction

Speaker recognition robustness in adverse condition has been investigated widely in recent years [1, 2, 3, 4, 5, 6]. There are quite a number of factors affecting the automatic speaker recognition performance including channel/session variability and noise/reverberation. In real-world applications dealing with mismatched condition is inevitable and any type of mismatch between training and test session will potentially result in degraded performance. Based on the type of the data in national institute of standards and technology (NIST) speaker recognition evaluations [7], the researchers in speaker recognition field have successfully developed techniques to deal with session/channel variability [1, 2, 3].

Although the state-of-the-art algorithms' sensitivity to unseen channel or session variability are partially mitigated, they are highly vulnerable to additive noise and reverberant environment [8, 5]. It has also been shown that even the performance of the state-of-the-art speaker recognition systems degrades substantially when limited speech is available in testing phase [9]. Although there are recent studies to handle reverberation and additive noise in feature [6] and model domain [4, 5] for speaker recognition systems, the compensation techniques with respect to noise and reverberation for speaker recognition systems are still an open question.

Multi-condition training entails including noisy samples of original data in the training phase to have parallel models for each speaker in different noise/SNR conditions. This technique has been shown to be an effective way of handling noisy condition specially when the expected noise type in test phase has already been observed in training phase [10, 4]. Training the speaker models with multi-conditioned noisy speaker models by incorporating missing feature principals has been studied for GMM-UBM based speaker recognition which provides considerable performance improvement over the models only trained with clean data [4]. GMM-UBM system has also shown an average 85% identification accuracy on GRID corpus (speech+speech mixture) when a mixed-UBM and multi-conditioned GMMs are utilized [11]. Multi-condition training for Gaussian PLDA-based speaker recognition on a subset of NIST SRE'10 interview data is shown to be an effective approach to handle additive noise and reverberation [5].

There are many speech enhancement algorithms proposed for robust automatic speech recognition (ASR), most of them relying on the assumption that the additive noise is a stationary process which is not always true for real-world applications. Minimum statistics [12], improved minima controlled recursive averaging [13], MMSE spectral amplitude estimator [14] and log-spectral amplitude estimator [15] are examples of these algorithms which essentially fail in non-stationary noise tracking and produce undesirable artifacts yet to be captured by recognition system during training phase [16]. Although there are more robust speech enhancement algorithms proposed to handle non-stationary reverberant scenarios [17, 18, 16], it is outside of the focus of this paper to investigate the effect of speech enhancement algorithms on speaker recognition performance.

Sparse representations have recently gained attention in speaker recognition [19, 20, 21, 22, 23]. An over-complete dictionary of speakers' GMM mean supervectors [24] can be utilized and then by finding the sparse activations for a test utterance, the identity inference build upon the activations [19, 20, 21]. Promising results have been reported by representing an utterance with its i-vector [3] instead of GMM mean supervector in an over-complete dictionary for speaker verification [22]. The i-vectors can also be computed subject to sparsity [23]. Exemplar-based sparse representation of speech signal has recently been found also useful in speech enhancement and speech recognition [25, 26].

In this work, we investigate whether an appropriate sparse representation can be used in the task of speaker recognition to improve robustness in dealing with additive noise. Specifically, we propose exemplar-based sparse representation, which can satisfactorily handle additive noise in speech recognition [26], for a closed-set speaker identification task. The difference with existing works in speaker recognition is that the dictionary is made by utilizing mel-frequency band energies in a large spectro-temporal context (25 frames) called exemplars. The dictionary elements are selected in such a way to be repre-

sentative of specific acoustic events. A set of noise exemplars is also included to allow coping with additive noises. The sparse activations are estimated to minimize the reconstruction error for the observation. In this work the activations are averaged over the utterance to make the secondary-feature for classification. Next these secondary-features are mapped onto sparse discriminant directions [27] followed by a dot-scoring for classification. The contribution of the current study is to introduce the exemplar-based sparse representation framework for speaker recognition and utilize sparse discriminant analysis to find speaker discriminant directions.

2. Sparse Representation and Discrimination

Speech and speaker recognition are two different tasks, still sharing many of their challenges and potential solutions. The fundamental problem is to construct a model of speech, and to match the observation to the model. In speech recognition, the speaking and pronunciation styles of different speakers can vary significantly. Therefore it has been found beneficial to train models for each speaker individually whenever possible. Meanwhile, the speaker-dependent models can also be used for speaker recognition by detecting the model best matching to the observed speech.

Recently, it has been demonstrated how speech can be modeled as a linear combination of long spectro-temporal segments, also known as speech *atoms* [25]. In the context of speech and speaker recognition, the sparse model based on spectro-temporal atoms has two major benefits. First, its inherent capability to model additive noise makes it suitable for recognition in adverse environments. Second, given enough atoms from multiple speakers, it can often capture both the phonetic content and the speaker identity of observations. The long temporal context of atoms is able to capture spectro-temporal patterns characteristic to words and speakers, which would not be possible using only momentary frame spectra.

It has been shown previously how to employ speech and noise atoms sampled directly from training material or the context to perform noise robust speech recognition [25, 28]. Atoms acquired this way are known as *exemplars*, and the approach as a whole is dubbed *exemplar-based sparse classification*. Whereas earlier work has mostly focused on recognizing speech using a single speaker's dictionary at a time, here the same approach is used with an important difference that we maintain a basis of all speakers' exemplars and utilize this *dictionary* in factorizing the observed utterance. Ideally, the exemplars from the speaker that originates the observed utterance will be activated.

2.1. Exemplar-based sparse representation

Let us denote the spectral magnitudes calculated in B mel-frequency bands in a sequence of T frames and reshaped into a vector, by $\mathbf{y}$. Fig. 1 shows the formation of a speech observation vector $\mathbf{y}$ for a single window. We can represent noisy speech observations as an additive combination of speech and noise atoms. Given an *exemplar dictionary* or *basis* of similarly vectorized speech and noise atoms $\mathbf{a}_j^s$ ($j = 1, \ldots, J$) and $\mathbf{a}_k^n$ ($k = 1, \ldots, K$), we can model the observation as

$$\mathbf{y} \approx \sum_{j=1}^{J} \mathbf{a}_j^s x_j^s + \sum_{k=1}^{K} \mathbf{a}_k^n x_k^n, \qquad (1)$$

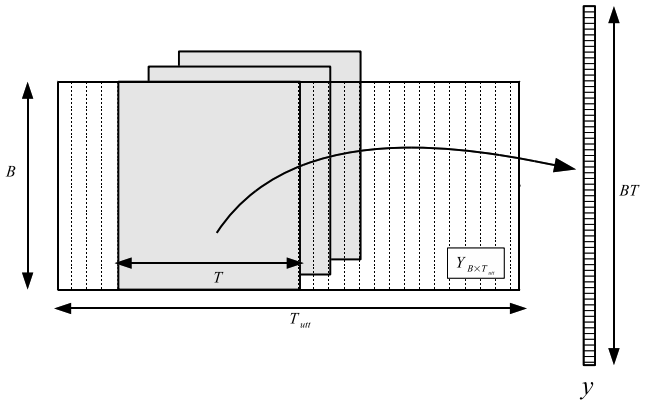

Figure 1: The formation of observation vectors: After converting an utterance to spectrogram representation $\mathbf{Y}$ with T_{utt} spectral amplitudes mapped to B mel-bands energies, a window of consecutive T frames are concatenated to form the observation window vector $\mathbf{y}$.

where the scalars x_j^s and x_k^n define the *activation weights* for each speech and noise atom, respectively ($L = J + K$ is the total number of exemplars). If we concatenate all the atom vectors as columns of a *basis matrix* $\mathbf{A}$ and the activations into a vector $\mathbf{x}$, the same model takes a matrix form $\mathbf{y} \approx \mathbf{A}\mathbf{x}$.

By finding a close approximation to actual $\mathbf{y}$ while minimizing the number of active atoms, we can construct a *sparse representation*, which often manages to reveal the most likely atomic components contributing to the observed mixture. If the signals are modeled in a domain which can be considered additive, it is also beneficial to enforce a *non-negativity* constraint on the weights. Consequently, the problem of finding either basis vectors $\mathbf{a}$, activations $\mathbf{x}$, or both is known as *non-negative matrix factorization* and the modeling technique is often referred to as *sparse coding* [29].

2.2. Convolutive spectral factorization

The model given in Equation (1) produces a length L activation vector for a single observation window. As utterances are generally longer than window length T, the whole observation spectrogram $\mathbf{Y}$ ($B \times T_{\text{utt}}$) is modeled in $W = T_{\text{utt}} - T + 1$ overlapping windows with a step of one frame. In earlier work, two different methods have been presented for handling the temporal continuity [26, 28]. Both produce an $L \times W$ *activation matrix* $\mathbf{X}$, each its columns representing activations in a single observation window.

In this work we use an algorithm referred to as *non-negative matrix deconvolution* (NMD), where all activations are used jointly for estimating the utterance spectrogram. The estimated spectrogram $\hat{\mathbf{Y}}$ is modeled convolutively as

$$\hat{\mathbf{Y}} = \sum_{t=1}^{T} \mathbf{A}_t \overrightarrow{\mathbf{X}}^{(t-1)}, \qquad (2)$$

where each $\mathbf{A}_t$ is a $B \times L$ matrix containing the t^{th} frame of all atom spectrograms, and $\overrightarrow{(\cdot)}$ shifts columns right within a $L \times T_{\text{utt}}$ matrix. The cost function to be minimized consists of Kullback-Leibler divergence for spectral distance between the observation and its estimate, and weighted $l1$-norm penalty for nonzero activations to enforce sparsity in the solution. The it-

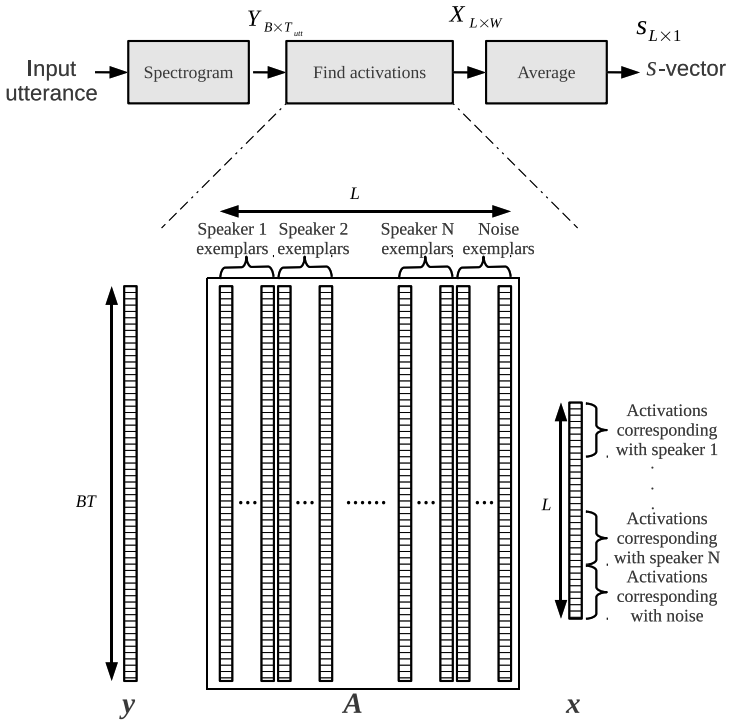

Figure 2: Forming a sparse identity s-vector from an input utterance. Every observation window activation is calculated with subject to sparsity to have $\mathbf{y} \approx \mathbf{A}\mathbf{x}$. The activations $\mathbf{x}$ are averaged over time to make s-vector.

erative update rules used to acquire $\mathbf{X}$ and other details of the factorization procedure can be found in [28, 30]. The activation matrix is averaged over window indices to form an s-vector ($L \times 1$) representing the atoms activated in the utterance. A schematic diagram of the process of converting an utterance to a fixed-dimension sparse vector $\mathbf{s}$ is shown in Fig. 2.

2.3. Sparse Discriminant Directions

It has been found very useful to classify low-dimensional i-vectors with LDA, probabilistic LDA [31, 32], and recently source normalized LDA [33]. However, it is known that when the number of predictor variables (feature dimension d) is much higher than the number of observations N, the conventional LDA algorithm fails [34]. Since the s-vector sparse representation is high-dimensional (d is number of exemplars here) we need to find an appropriate way of classification. Probabilistic LDA works on the assumption of Gaussian distribution for input vectors and models the data generation by the summation of speaker-dependent term, $\mu + \mathbf{F}\mathbf{h}_i$ considering I speakers and an utterance dependent term $\mathbf{G}\boldsymbol{w}_{im} + \epsilon_{im}$ with M utterances for speaker i [31]. The overall mean of the training vectors is denoted by μ and the matrices $\mathbf{F}$ and $\mathbf{G}$ are composed of basis for between-speaker and within-speaker subspaces, respectively. The $\mathbf{h}_i$ and $\boldsymbol{w}_{im}$ are positioning the input vector in between-speaker and within-speaker subspaces, respectively and ϵ_{im} is a Gaussian residual error term.

The LDA projection matrix is formed by first making an eigen-analysis on between-speaker ($\mathbf{S}_b$) and within-speaker ($\mathbf{S}_w$) covariance matrices and then using a subset of eigenvectors, $\boldsymbol{\upsilon}_j$ having the largest eigenvalues. Considering K speakers, the eigenvectors in LDA are found using the Fisher's criterion to find directions that maximize the between-speaker separation while keeping the within-speaker variation small;

$$\arg\max_{\boldsymbol{\upsilon}_j} \boldsymbol{\upsilon}_j^T \mathbf{S}_b \boldsymbol{\upsilon}_j \tag{3}$$

with subject to orthogonality constraint

$$\boldsymbol{\upsilon}_l^T \mathbf{S}_w \boldsymbol{\upsilon}_j = \begin{cases} 0 & l \neq j \\ 1 & l = j \end{cases}, j = 1, \ldots, K-1 \tag{4}$$

Penalized discriminant analysis (PDA) was proposed in [34] to account for $d \gg N$ condition. Sparse discriminant analysis (SDA) is recently proposed with the same principals of PDA to work with high dimensional sparse data while providing sparse eigenvectors in the solution [27]. The SDA essentially entails penalizing $\mathbf{S}_w$ to be $\mathbf{S}_w + \lambda_2\Omega$ where Ω is a penalty function introduced in [34] to take care of $d \gg N$ condition and changing the eigenvector selection criterion as

$$\arg\max_{\boldsymbol{\upsilon}_j} \boldsymbol{\upsilon}_j^T \mathbf{S}_b \boldsymbol{\upsilon}_j - \lambda_1 ||\boldsymbol{\upsilon}_j||_1 \tag{5}$$

to produce eigenvectors with p non-zero values [27]. λ_1 and λ_2 are optimization parameters and the optimization to find the sparse discriminant directions is performed with *elastic net* [35]. The elastic net is particularly favored over *lasso*-based optimization [36] in case when the feature dimension is much bigger than the number of observations.

3. Experiments

We perform text-constrained closed-set speaker identification to evaluate the performance of the proposed system. HMM-based text-dependent speaker recognition and GMM-UBM based text-independent speaker recognition are employed as the baseline methods for performance comparison. In addition to simple manipulation of exemplar activations, we employ PLDA and SDA to model the sparse exemplar activations.

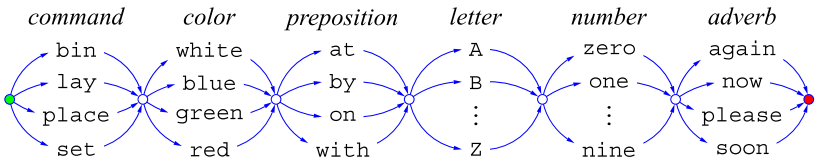

Figure 3: CHiME corpus sentences grammatical structure

3.1. Database description

For performance evaluation, we conducted our experiments on the PASCAL *computational hearing in multi source environments* (CHiME) speech separation and recognition challenge dataset [37]. The CHiME evaluation data is derived from convolving the clean speech signals (extracted from the GRID audio-visual corpus [38]) with real room impulse response to simulate the reverberant environment as well as adding wide range of noise sources collected from a living room at different locations. The GRID corpus consists of 34,000 distinct utterances from 34 speakers (18 males and 16 females) and as it is shown in Fig. 3 the sentences follow a unique grammatical structure each composed of a combination of as six word commands: verb, color, preposition, letter, digit and coda (e.g., "*bin white at p nine soon*"). The keywords emphasized for speech intelligibility or recognition task in challenge are the items in position 4 and 5 referring to letter and digit, respectively. The possible letters are 25 English alphabet letters and finally the digits are selected from 0 to 9.

The CHiME dataset is sub-divided into three parts: training, development and test sets. For each speaker, 500 clean (reverberated) utterances are provided for training purposes. Each of development and test sets are composed of 600 utterances mixed at six SNR levels ranging from -6dB to 9dB. The noise contamination was done by challenge organizers which the details can be found in [37]. The SNR levels encountered in development and test sets are expected to be representative of real-world situation except the fact that it is not accounting for the *Lombard effect*. It should be noted that noise types are different across different SNR-levels. The sentences were originally stereophonic and sampled at 48kHz and provided also in 16kHz format. The recognizers in this work use single-channel signals by averaging the two channels together in waveforms and we use 16 kHz data in our experiments. Summing two channels is equivalent to a delay-and-sum beamformer where no delay estimation is employed. All the analysis is performed on reverberated speech as a straightforward approach for considering the effect of reverberation in modeling.

3.2. Experimental protocol

The training set is used to train speaker models, while the development and test sets are used to report the system performance in terms of the speech enhancement, speech recognition, and in this paper speaker recognition accuracy. Here we are using the speakers' training material to build up speaker models. The development set is used to tune the PLDA and SDA parameters. The tuned system is then tested on unseen test set to evaluate the generalization capability of trained model. The criterion for optimizing the system performance on development set is considered as the identification accuracy over all SNR ranges in development set. This ensures us that the recognizer parameters are not optimized for a specific condition.

Table 1: ASR results for key-word recognition accuracy in percent on unprocessed mixture (no speech enhancement applied) on CHiME corpus.

	SNRs					
	9dB	6dB	3dB	0dB	-3dB	-6dB
Development set	83.1	73.8	64.0	49.1	36.7	31.1
Test set	83.8	74.3	62.4	48.3	37.4	31.6

3.3. Text-Dependent HMM

Since the setup of the CHiME corpus is text constrained, the first baseline approach is selected to be a text-dependent HMM approach where basically speaker-dependent HMMs are trained for speech recognition. We take the baseline recognizer supplied by the CHiME challenge organizers [37] and instead of decoding the observed utterance with only the known speaker model, we let all the speaker-dependent HMMs decode the observation and identify the speaker as the one who provides highest log-likelihood. The words are modeled as whole-word HMMs with a left-to-right model topology with no skips over states and 7 Gaussian mixtures per state with diagonal covariance matrices [37]. Considering 2 states per phoneme the number of states for each word is given as:

4 states: at by in a b c d e f g h i j k l m n o p q r s t u v x y z one two three eight

6 states: bin lay place set blue green red white with four five six nine now please soon

8 states: again zero

10 states: seven

This leads to overall number of 250 states in the HMM structure. Speaker-dependent HMMs are trained by first estimating a set of speaker independent HMMs as the starting point, and then performing 4 more iterations of EM training using the 500 training utterances for each speaker. The HMMs are trained using reverberant signals without any noise and there is neither adaptation to noisy signals nor multi-conditional training. For HTK the training and test scripts provided for the CHiME challenge were used [37]. The features are cepstral mean normalized MFCCs which include 12 base coefficients plus energy, concatenated with delta and acceleration features.

We report the ASR results averaged on development set in Table 1 to show the key-word recognition performance of the baseline HMM for detecting color and letter.

3.4. Text-Independent GMM-UBM

Although the speaker recognition task in context of CHiME corpus is a *text-constrained*, in addition to HMM-based approach we have also evaluated a baseline text-independent GMM-UBM approach [39] as our second benchmark method. To this end, using the same MFCC features as HMM-based approach we pooled all the target speakers training data to

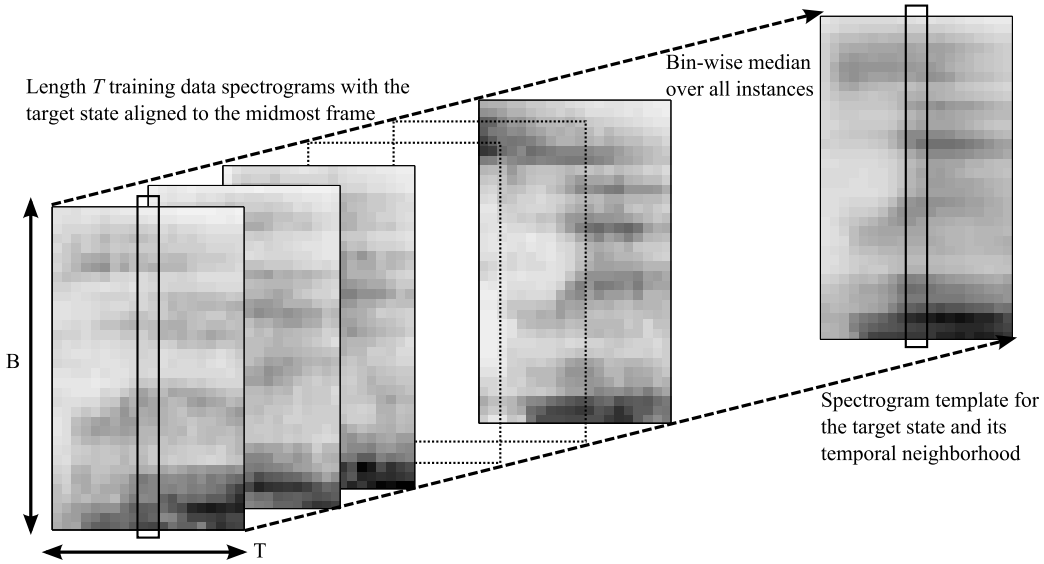

Figure 4: Forming an atom template for a single phonetic state. Training data spectrogram segments, where the target state appears, are placed in a $B \times T$ window with the target state in the middle. A bin-wise median is taken over instances to model the state spectrum and its context with a single template.

make a 512 Gaussian UBM and then use each speakers 500 training segments to make a *maximum a posteriori* estimation for each speaker's GMM model.

3.5. Exemplar-based approach

Setting up the exemplar-based system, all the audio was converted into spectral magnitudes in 40 mel bands (B) at a temporal resolution of 25 ms frame length and 10 ms frame shift. The observation window length T was set to 25 frames (265 ms). A 250-atom speech basis was generated for each speaker by constructing spectrogram templates for the 250 speech states in the system.

For each state in turn, the spectrograms of all instances of the speaker's training utterances containing the target state were gathered together. Using forced alignment information from the CHiME HTK models (as described in section 3.3), spectrograms of the target state and its immediate neighborhood were placed in a $B \times T$ window with the target state in the middle. Thereafter a bin-wise median was taken over all instances to construct a spectro-temporal template of the state and its typical context. The process is visualized in Figure 4.

By repeating the procedure for all states and speakers, an 8500-atom (250 atoms $\times$ 34 speakers) combined speech basis was acquired. The speaker-specific atoms were estimated using 300 out of 500 available files in training set per speaker. In addition, 250 noise atoms were extracted adaptively from the local noise context of the utterance to be recognized as in [26]. All in all, the speech and noise atoms formed an $L = 8750$ atom dictionary. The spectral bands of both basis and utterance features were re-weighted using a band-normalizing curve acquired from speech training material. Individual atoms were normalized to unitary Euclidean norm over their whole spectrogram content. Thereafter the development and test utterances were factorised as described in Section 2.2, and $\mathbf{s}$-vectors were averaged from the activation matrices.

Four speaker identification system are built based on employing $\mathbf{s}$-vectors as their input for classification. The speaker identification error rates for these four approaches along with two baseline systems are presented in Table 2. Since the activation data are sparse, conventional LDA cannot be applied directly to classify them. Extracting 200 sparse representation $\mathbf{s}$-vectors for the remaining 200 training files per speaker, we first clip them to have only 8500 speaker-specific activations, then length normalize and finally employ PLDA and SDA to find the most discriminant directions.

Exemplar-based simple manipulation In simple manipulation of the activation vectors, since we already know which activations are corresponding to speakers' exemplars in dictionary, we can average over activations for the exemplars for a specific speaker to get the first no-modeling score for each speaker.

Exemplar-based + dot-scoring We utilize the speaker-specific exemplars and by averaging them make an average sparse representation for each speaker. Applying a simple inner product $\langle \mathbf{s}_1, \mathbf{s}_2 \rangle$, dubbed as "dot-scoring", between speakers' average sparse representation and test $\mathbf{s}$-vector will generate the recognition score.

Exemplar-based PLDA The PLDA dimensions optimizing the overall performance on development set was found to be 33 eigen-voices and 1 eigen-channels. This is an interesting observation because increasing the number of eigen-channels would improve the recognition on clean and slightly noisy conditions but ruins out the noisy condition performance.

Exemplar-based + SDA + dot-scoring After projecting the 200 $\mathbf{s}$-vectors on sparse directions for each speaker, we take the average of them to present the speaker identity vector. A recognition score is defined as a dot-scoring of projected test $\mathbf{s}$-vector and speakers' identity vectors. The penalty function Ω for penalizing the within-class covaraince matrix in SDA is set to identity matrix which

Table 2: Speaker identification error rate (in percent) comparison of different systems. The identification errors are measured on CHiME corpus development and test set for each SNR utilizing 600 utterances each of average 2 seconds length.

	Development						
SNR	Clean	9dB	6dB	3dB	0dB	−3dB	−6dB
HMM	0.3	1.7	6.7	17.8	33.8	51.8	62.8
GMM-UBM	0.3	4.7	6.7	15.8	23.8	30.2	37.7
Exemplar-based simple manipulation	0.5	1.2	2.0	2.7	8.3	16.5	32.7
Exemplar-based + dot-scoring	1.2	1.5	1.5	4.0	7.8	12.0	23.3
Exemplar-based PLDA	0.2	0.7	1.0	2.3	7.7	16.0	34.2
Exemplar-based + SDA + dot-scoring	0.3	0.5	0.7	1.7	3.2	8.2	17.3
	Test						
HMM		2.0	6.8	16.8	41.2	56.7	67.3
GMM-UBM	-	2.2	8.2	14.7	26.3	35.3	43.7
Exemplar-based simple manipulation	-	1.0	2.0	3.7	7.7	14.8	34.8
Exemplar-based + dot-scoring	-	1.7	2.5	3.5	7.0	10.3	22.7
Exemplar-based PLDA	-	0.5	0.8	2.3	8.3	14.5	33.7
Exemplar-based + SDA + dot-scoring	-	0.2	0.7	1.3	4.5	5.7	17.5

corresponds to a *ridge regression* scenario. The parameter λ_2 is set to 0.01 meaning that by adding a small number to the diagonal of the wihtin-class covariance matrix it gets full rank and that takes care of the $d \gg N$ condition. We have optimized on whole development set the number of non-zero elements for sparse discriminant direction in SDA by fixing the number of discriminant directions to be 33. Te optimal number of non-zero elements was found to be 500.

3.6. Results and Discussion

The speaker identification performance of baseline systems along with proposed exemplar-based approach are presented in Table 2. It is observed that baseline systems only show a reasonable performance on the clean and slightly noisy (SNR$=$ 9dB) conditions and the performance drops substantially for SNR$<$ 9dB. The GMM-UBM system shows more robust behavior in low SNRs compared to text-dependent HMM system. The HMM-based system performance decline in noisy conditions for speaker identification is much bigger than for speech recognition. It yet remains to be investigated if applying speech enhancement algorithms, which most of the time increase the ASR performance, will translate to improved speaker recognition accuracy.

The simple manipulation of $\mathbf{s}$-vectors in development and test set (without any training) provides much better speaker identification results than baseline systems for SNR$\leq$ 9dB. This great improvement is a result of inherent capability of exemplar-based approach to capture long spectro-temporal context and careful design of the system which compresses speaker-specific information into $\mathbf{s}$-vectors. Although the exemplar-based system utilizes the noise exemplars in estimating the activations (baseline HMM and GMM-UBM system do not incorporate noise information in modeling), the small performance difference between development and test sets reveals that this representation is indeed robust to unseen noise conditions. Employing dot-scoring directly on $\mathbf{s}$-vectors to calculate the recognition score provides reduced error in SNR$\leq$ 0dB. This phenomenon can be described as the noisy speech can activate the exemplars in a different way than they are activated in clean speech and hence by using average $\mathbf{s}$-vector of each speaker, the more relevant activations are emphasized in dot-scoring.

By employing a modeling approach on $\mathbf{s}$-vector domain both PLDA and SDA bring extra performance improvement over simple manipulation of $\mathbf{s}$-vectors. The PLDA essentially helps in SNR$>$ 0dB which is a result of optimizing the discriminative directions to the clean data. The performance gain in the order of magnitude for SDA-based approach in SNR$\leq$ 0dB enlightens the suitability of sparse discriminative directions for projecting the $\mathbf{s}$-vectors. Employing McNemars' statistical test, it is found out that all the results for SNR$\leq$ 6dB for exemplar-based approach are significantly different than of baseline systems. It is again a topic for further research if finding sparse discriminant directions is only helpful in handling sparse data or can be useful also in dealing with information rich i-vectors.

4. Conclusions

A new approach for closed-set speaker identification based on exemplar-based representation and sparse discrimination is proposed. Evaluating on recently developed CHiME corpus we have found the proposed system outperforming the baseline GMM-UBM and HMM based systems with a large margin.

5. Acknowledgment

The work of Rahim Saeidi was funded by the European Community's Seventh Framework Program (FP7/2007-2013) under grant agreement no. 238803. Authors would like thank Dr. Mitchell McLaren for fruitful discussions and Dr. Line Clemmensen for sharing her implementation of sparse discriminant analysis.

6. References

[1] A. Solomonoff, W.M. Campbell, and I. Boardman. Advances in channel compensation for SVM speaker recognition. In *Proc. Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP 2005)*, pages 629–632, Philadelphia, USA, March 2005.

[2] P. Kenny, P. Ouellet, N. Dehak, V. Gupta, and P. Dumouchel. A study of interspeaker variability in speaker verification. *IEEE Trans. Audio, Speech and Language Processing*, 16(5):980 –988, july 2008.

[3] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet. Front-end factor analysis for speaker verification. *IEEE Trans. Audio, Speech and Language Processing*, 19(4):788 –798, may 2011.

[4] J. Ming, T.J. Hazen, J.R. Glass, and D.A. Reynolds. Robust speaker recognition in noisy conditions. *IEEE Trans.*

Audio, Speech and Language Processing, 15(5):1711 – 1723, july 2007.

[5] D. Garcia-Romero, X. Zhou, and C. Y. Espy-Wilson. Multicondition training of gaussian plda models in i-vector space for noise and reverberation robust speaker recognition. In *Proc. Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP 2012)*, 2012.

[6] C.M. Vannicola, B.Y. Smolenski, B. Battles, and P.A. Ardis. Mitigation of reverberation on speaker identification via homomorphic filtering of the linear prediction residual. In *Proc. Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP 2011)*, pages 5512–5515, may 2011.

[7] NIST speaker recognition evaluation.

[8] M. I. Mandasari, M. McLaren, and D. A. van Leeuwen. The effect of noise on modern automatic speaker recognition systems. In *Proc. Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP 2012)*, 2012.

[9] M. I. Mandasari, M. McLaren, and D. A. van Leeuwen. Evaluation of i-vector speaker recognition systems for forensic application. In *Proc. Interspeech 2011*, pages 21–24, 2011.

[10] R. Lippmann, E. Martin, and D. Paul. Multi-style training for robust isolated-word speech recognition. In *Proc. Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP 1987)*, volume 12, pages 705 – 708, apr 1987.

[11] R. Saeidi, P. Mowlaee, T. Kinnunen, Z. H. Tan, M. G. Christensen, P. Fränti, and S. H. Jensen. Signal-to-signal ratio independent speaker identification for co-channel speech signals. In *Proc. IEEE Int. Conf. Pattern Recognition (ICPR 2010)*, pages 4545–4548, 2010.

[12] R. Martin. Noise power spectral density estimation based on optimal smoothing and minimum statistics. *IEEE Trans. on Speech and Audio Processing*, 9(5):504 –512, jul 2001.

[13] I. Cohen. Noise spectrum estimation in adverse environments: improved minima controlled recursive averaging. *IEEE Trans. on Speech and Audio Processing*, 11(5):466 – 475, sept. 2003.

[14] Y. Ephraim and D. Malah. Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator. *IEEE Trans. Acoustics, Speech, and Signal Processing*, 32(6):1109 – 1121, dec 1984.

[15] Y. Ephraim and D. Malah. Speech enhancement using a minimum mean-square error log-spectral amplitude estimator. *IEEE Trans. Acoustics, Speech, and Signal Processing*, 33(2):443 – 445, apr 1985.

[16] P. Mowlaee, R. Saeidi, and R. Martin. Model-driven speech enhancement for multisource reverberant environment (signal separation evaluation campaign sisec 2011). In *Proc. 10th International Conference on Latent Variable Analysis and Source Separation (LVA/ICA 2012)*, 2012.

[17] J.S. Erkelens and R. Heusdens. Tracking of nonstationary noise based on data-driven recursive noise power estimation. *IEEE Trans. Audio, Speech and Language Processing*, 16(6):1112 –1123, aug. 2008.

[18] R. C. Hendriks, R. Heusdens, and J. Jensen. Mmse based noise psd tracking with low complexity. In *Proc. Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP 2010)*, pages 4266 –4269, march 2010.

[19] I. Naseem, R. Togneri, and M. Bennamoun. Sparse representation for speaker identification. In *Proc. IEEE Int. Conf. Pattern Recognition (ICPR 2010)*, pages 4460 – 4463, aug. 2010.

[20] J.M.K. Kua, E. Ambikairajah, J. Epps, and R. Togneri. Speaker verification using sparse representation classification. In *Proc. Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP 2011)*, pages 4548–4551, may 2011.

[21] M. Li and S. Narayanan. Robust talking face video verification using joint factor analysis and sparse representation on gmm mean shifted supervectors. In *Proc. Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP 2011)*, May 2011.

[22] M. Li, X. Zhang, Y. Yan, and S. Narayanan. Speaker verification using sparse representations on total variability i-vectors. In *Proc. Interspeech 2011*, August 2011.

[23] M. Li, S. Narayanan, C. Lu, and A. Wang. Speaker verification using lasso based sparse total variability supervector and probabilistic linear discriminant analysis. In *NIST 2011 Speaker Recognition Workshop*, December 2011.

[24] W.M. Campbell, D.E. Sturim, and D.A. Reynolds. Support vector machines using GMM supervectors for speaker verification. *IEEE Signal Processing Letters*, 13(5):308–311, May 2006.

[25] J. F. Gemmeke, T. Virtanen, and A. Hurmalainen. Exemplar-based sparse representations for noise robust automatic speech recognition. *IEEE Trans. Audio, Speech and Language Processing*, 19(7):2067 – 2080, 2011.

[26] A. Hurmalainen, K. Mahkonen, J. F. Gemmeke, and T. Virtanen. Exemplar-based recognition of speech in highly variable noise. In *International Workshop on Machine Listening in Multisource Environments*, pages 1–5, march 2011.

[27] L. Clemmensen, T. Hastie, D. Witten, and B. Ersboll. Sparse discriminant analysis. *Technometrics*, 54(4):406–413, 2011.

[28] A. Hurmalainen, J. Gemmeke, and T. Virtanen. Non-negative matrix deconvolution in noise robust speech recognition. In *Proc. Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP 2011)*, pages 4588 – 4591, 2011.

[29] T. Virtanen. Separation of sound sources by convolutive sparse coding. In *in Proceedings of ISCA Tutorial and Research Workshop on Statistical and Perceptual Audio Processing*, 2004.

[30] T. Virtanen. *Sound source separation in monaural music signals*. PhD thesis, Tampere University of Technology, Tampere, Finland, 2006.

[31] S. J. D. Prince and J. H. Elder. Probabilistic linear discriminant analysis for inferences about identity. In *11th International Conference on Computer Vision*, pages 1–8, 2007.

[32] P. Kenny. Bayesian speaker verification with heavy-tailed priors. In *Proc. IEEE Odyssey: the Speaker and Language Recognition Workshop (Odyssey 2010)*, Brno, Czech Republic, June 2010.

[33] M. McLaren and D. van Leeuwen. Source-normalized lda for robust speaker recognition using i-vectors from multiple speech sources. *IEEE Trans. Audio, Speech and Language Processing*, 20(3):755–766, march 2012.

[34] T. Hastie, A. Buja, and R. Tibshirani. Penalized discriminant analysis. *Annals of Statistics*, 23(1):73–102, 1995.

[35] H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B*, 67(2):301–320, 2005.

[36] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B*, 58(1):267–288, 1996.

[37] H. Christensen, J. Barker, N. Ma, and P. Green. The CHiME corpus: a resource and a challenge for computational hearing in multisource environments. In *Proc. Interspeech*, pages 1918–1921, 2010.

[38] M. Cooke, J. R. Hershey, and S. J. Rennie. Monaural speech separation and recognition challenge. *Elsevier Computer Speech and Language*, 24(1):1–15, 2010.

[39] D.A. Reynolds, T.F. Quatieri, and R.B. Dunn. Speaker verification using adapted gaussian mixture models. *Digital Signal Processing*, 10(1):19–41, January 2000.