ELSEVIER

Available online at www.sciencedirect.com

SciVerse ScienceDirect

Speech Communication 55 (2013) 769–781

SPEECH COMMUNICATION

www.elsevier.com/locate/specom

In-set/out-of-set speaker recognition in sustained acoustic scenarios using sparse data

John H.L. Hansen *, Jun-Won Suh, Matthew R. Leonard [1]

Center for Robust Speech Systems (CRSS), Department of Electrical Engineering, Erik Jonsson School of Engineering and Computer Science, University of Texas at Dallas, Richardson, TX 75080-1407, USA

Received 25 November 2011; received in revised form 21 January 2013; accepted 24 January 2013
Available online 27 February 2013

Abstract

This study addresses the problem of identifying in-set versus out-of-set speakers in noise for limited train/test durations in situations where rapid detection and tracking is required. The objective is to form a decision as to whether the current input speaker is accepted as a member of an enrolled in-set group or rejected as an outside speaker. A new scoring algorithm that combines log likelihood scores across an energy-frequency grid is developed where high-energy speaker dependent frames are fused with weighted scores from low-energy noise dependent frames. By leveraging the balance between the speaker versus background noise environment, it is possible to realize an improvement in overall equal error rate performance. Using speakers from the TIMIT database with 5 s of train and 2 s of test, the average optimum relative EER performance improvement for the proposed full selective leveraging approach is +31.6%. The optimum relative EER performance improvement using 10 s of NIST SRE-2008 is +10.8% using the proposed approach. The results confirm that for situations in which the background environment type remains constant between train and test, an in-set/out-of-set speaker recognition system that takes advantage of information gathered from the environmental noise can be formulated which realizes significant improvement when only extremely limited amounts of train/test data is available.

Keywords: Speaker recognition; In-set/out-of-set; Sparse data; Environmental noise

1. Introduction

In-set/out-of-set speaker recognition systems are useful for situations where it is important to detect and track the presence of speakers in a group. Examples of speech systems that benefit from in-set/out-of-set recognition include dialog systems, communications systems, spoken document retrieval, and security applications that allow access of private information for only people belonging to a specific group of authorized users (Angkititrakul and Hansen, 2007; Hansen et al., 2005; Prakash and Hansen, 2007; Suh and Hansen, 2012). The objective of an in-set/out-of-set speaker recognition system is to make a decision as to whether to accept the claim that the current input speaker is a legitimate member of the enrolled in-set group, or to reject the claim and classify the speaker as an outside speaker.

There are two aspects related to the formulation of the problem in this study. First, the ability to leverage speaker versus environment knowledge is employed, and for the second, the ability to determine if an input speaker is a member of an "in-set" versus "out-of-set" group. It is important to differentiate this problem from other research

* Corresponding author. Address: Center for Robust Speech Systems (CRSS), Erik Jonsson School of Engineering and Computer Science, Dept. of Electrical Engineering, University of Texas at Dallas, 2601 N. Floyd Road, EC33, Richardson, TX 75080-1407, USA. Tel.: +1 972 883 2910; fax: +1 972 883 2710.

E-mail address: John.Hansen@utdallas.edu (J.H.L. Hansen).

URL: http://crss.utdallas.edu (J.H.L. Hansen).

[1] This project was funded by AFRL through a subcontract to RADC Inc. under FA8750-09-C-0067, and partially by the University of Texas at Dallas from the Distinguished University Chair in Telecommunications Engineering held by J.H.L. Hansen.

0167-6393/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.specom.2013.01.006

tasks consider in the field for speaker recognition such as those considered in the NIST SRE (Speaker Recognition Evaluation). In this study, the following research assumptions are made:

- Only a limited amount of training data is available for each speaker (i.e., approximately 5 s of training data per speaker).
- It is not necessary to specifically identify who is speaker; only if the input speaker is part of the "in-set" group, and if not, set that speaker aside as being a member of the "out-of-set" group.
- Only a limited amount of test data is available for each input speaker (i.e., 2 s for any test set speaker).
- While all speaker recognition solutions capitalize on setting aside silence, noise and non-speech/speaker related input sample data, this study instead seeks to leverage the knowledge of the particular environment the speaker is in while producing speech. Since this study allows for only 5 s of train, and 2 s of test data, the task of In-Set/Out-of-Set Speaker ID is very challenging, and therefore we assume (i) there is no mismatch in terms of microphone or communications handset, and (ii) that the subject remains in the same acoustic environment between train and test. This second assumption is deemed both acceptable and common in cases where communications takes place between subjects communicating between each other in mobile environments and within a short time window (i.e., drivers in various consumer, commercial or military vehicles talking with each other).

Fig. 1 illustrates the two main issues for this study. In Fig. 1, the difference between speaker recognition and combined speaker/environment recognition is shown. Here, we choose to capitalize on the fact that the speaker is not likely to change vehicles in a mobile context within a short time window (i.e., within seconds or minutes from the audio stream), and therefore it is possible to reduce speaker identification (ID) confusion since the probability that confusable speakers are also in the same or confusable noise environments is lower. Therefore, when speakers are clearly separable for speaker ID, the in-set/out-of-set speaker recognition system could simply emphasize the log likelihood score of speaker ID. If there is uncertainty in the decision, combining the scores from both speaker space and environment space would improve performance. This is because, as shown in Fig. 1, that speakers are not as likely to be in the same environment space and be confusable at the same time. We recognize there are many other challenges associated with effective speaker recognition, and that these are all important. However, in this study, the focus is exclusively limited to (i) small train/test data sets where (ii) it is acceptable to take advantage of the acoustic environmental space in making a final output decision for in-set/out-of-set speaker recognition.

The area of speaker recognition has seen significant research efforts in recent years. The U.S. NIST SRE has resulted in numerous submissions from across the U.S. and the world (50 groups participated in 2010) (National Institute of Standards and Technology, 2009). In addition, effort has also been made in addressing robustness for speaker identification due to channel, noise, and speaker variability, including features (Shao et al., 2007), modeling (Rose et al., 1994), and normalization (Z-, T-, H-norm) (Ariyaeeinia et al., 2006; Auckenthaler et al., 2000). However, with respect to the NIST SRE, the tasks in 2006, 2008, 2010 have all focused on noise free data with only (i) handset, (ii) microphone, and (iii) communication channel variability. Speaker variability such as stress, task, etc. has not been consider nor has emotion. An effort was made

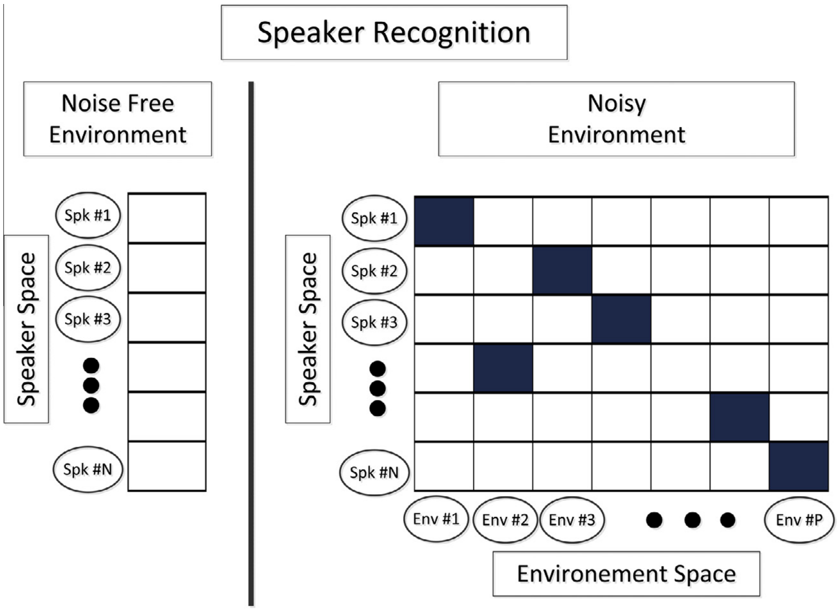

Fig. 1. Speaker recognition in noise free/noisy environment.

to consider high and low vocal effort in low, but this portion lacked the structure necessary to investigating of vocal effort whisper, soft, loud, shout.

The evaluation of in-set speaker recognition is based on two error measurements. The first, false rejection (FR), occurs when a member of the enrolled in-set group is rejected and classified as belonging to the out-of-set group; the second, false acceptance (FA), occurs when an outside speaker is accepted as being part of the in-set group. One of the main challenges for this type of system is effective rejection of outliers, while allowing speech production variability for the in-set speakers, such as interspeaker variations at the segmental level (Doddington, 1985). For this study, modeling is performed with what has become the dominant approach in text-independent speaker recognition: Gaussian Mixture Models (GMMs) with a UBM and maximum a posteriori (MAP) speaker adaptation (Prakash and Hansen, 2007; Reynolds et al., 2000; Xiang and Berger, 2003). Basic in-set/out-of-set speaker recognition is performed as follows: a speaker-independent universal background model (UBM) is generated from an available set of non-target speakers, using the expectation–maximization (EM) algorithm. For a speaker model Λ_n and D-dimensional observation vector $\boldsymbol{x}_t$, the probability density function (pdf) of an M-component Gaussian is:

$$p(\boldsymbol{x}_t|\Lambda_n) = \sum_{m=1}^{M} \boldsymbol{\omega}_{\boldsymbol{nm}} \mathcal{N}_{nm}(\boldsymbol{x}_t), \tag{1}$$

where $\boldsymbol{\omega}_{\boldsymbol{nm}}$ is the mixture weight of the mth component unimodal Gaussian density $\mathcal{N}_{nm}(\boldsymbol{x}_t)$. This Gaussian density is assumed to have a mean vector $\boldsymbol{\mu}_{nm}$ and diagonal covariance matrix $\boldsymbol{\Sigma}_{nm}$,

$$\mathcal{N}_{nm}(\boldsymbol{x}_t) = \frac{1}{(2\pi)^{\frac{D}{2}}|\boldsymbol{\Sigma}_{nm}|^{\frac{1}{2}}} e^{-\frac{1}{2}(\boldsymbol{x}_t-\boldsymbol{\mu}_{nm})^T \boldsymbol{\Sigma}_{nm}^{-1}(\boldsymbol{x}_t-\boldsymbol{\mu}_{nm})}. \tag{2}$$

Speaker dependent GMMs are then trained for all target in-set speakers by MAP adaptation of the UBM parameters $\boldsymbol{\omega}_{0m}, \boldsymbol{\mu}_{0m}, \boldsymbol{\Sigma}_{0m}$. This adaptation process is employed due to its ability to cope with extremely limited data ($\sim$5 s) (Prakash and Hansen, 2007). Based on experimental results, the best performance is achieved using only mean adaptation. The mean $\hat{\boldsymbol{\mu}}_{nm}$ of the mth component of Λ_n is updated via

$$\hat{\boldsymbol{\mu}}_{nm} = \frac{\eta_m}{\eta_m+\gamma} E_m(\boldsymbol{X}_n) + \frac{\gamma}{\eta_m+\gamma} \boldsymbol{\mu}_{nm}, \tag{3}$$

where γ is relevance factor, retained for all baseline settings in our experiments, that controls the adaptation balance between UBM parameters and training observations. Finally, η_m and $E_m(\boldsymbol{X}_n)$ can be computed as,

$$P(m|\boldsymbol{x}_{nt}) = \frac{\omega_{nm}\mathcal{N}_{nm}(\boldsymbol{x}_{nt})}{\sum_{j=1}^{M} \omega_{nj}\mathcal{N}_{nj}(\boldsymbol{x}_{nt})}, \tag{4}$$

$$\eta_m = \sum_{t=1}^{T_n} P(m|\boldsymbol{x}_{nt}), \tag{5}$$

$$E_m(\boldsymbol{X}_n) = \frac{1}{\eta_m} \sum_{t=1}^{T_n} P(m|\boldsymbol{x}_{nt}) \cdot \boldsymbol{x}_{nt}. \tag{6}$$

The speaker-dependent model obtained from the MAP-adapted UBM serves two purposes: (i) it provides a tighter coupling between the speaker specific models and the UBM, and (ii) it helps to mitigate the problem of sparseness that results from limited enrollment data.

When a test speaker is submitted to the system, two alternate Mel Frequency Cepstral Coefficient (MFCC) parameterizations are used, one for each of two databases: TIMIT (Garofolo et al., 1993) and NIST SRE-2008 (National Institute of Standards and Technology, 2008). For evaluations involving TIMIT, a 19-dimensional static MFCC feature is used with a frame size of 30 ms and skip rate of 10 ms. For evaluations involving NIST SRE data, 19-dimensional static and dynamic ($\triangle + \triangle\triangle$) MFCCs with log energy are used with a frame size of 25 ms and frame skip rate of 10 ms. During testing, the extracted MFCCs for the input speaker are tested against each of the in-set speaker models, as well as the UBM. The model with the highest probability is selected as the input speaker. Therefore, for an in-set size of N speakers, the input speaker has a total of $N+1$ possible classifications, consisting of any one of the N in-set speakers, or the UBM. If any of the in-set speakers is selected, the input speaker's claim is accepted; otherwise, the UBM has the highest probability and the claim is rejected. The fundamental idea explored here is that for extremely small train/test size data sets can knowledge of the background acoustic environment classification aid in improving in-set/out-of-set speaker recognition? Therefore, for this study, it is assumed and experimented that the speaker remains to the same environment.

This study is organized as follows. In Section 2, the corpus and background noise types used are discussed. Section 3 considers both noise and in-set/out-of-set speaker recognition. Section 4 presents the proposed new approach which incorporates models across an energy-frequency grid to obtain information from both high-energy speaker dependent frames and low-energy noise dependent frames, and reports on a series of experimental results for an initial version of the approach. Section 5 presents the foundation and results of the complete selective based speaker/environment leveraging system. Finally, conclusions and discussion for future work are presented in Section 6.

2. Corpus and noise types

In this study, evaluations are performed using both the TIMIT corpus and NIST SRE-2008 corpus. For the in-set/out-of-set scenario, there are 60 total speakers, divided into 15 in-set speakers and 45 out-of-set speakers. The short duration of train and test data is required for rapid enrollment and detection of speakers. The sparse data causes the acoustic mismatch in train and test data (Prakash and Hansen, 2007; Suh and Hansen, 2012), and this study focuses on sparse data problem in noisy background.

The sparse train (5 s) and test data (2 s) case are very specific sparse data problem, and the various data size of sparse train and test experiments were studied in previous study (Suh and Hansen, 2012).

For TIMIT evaluations, the training data is 5 s in duration and test data is 2 s in duration. System evaluations are also performed on the 5 min core condition of NIST SRE-2008. The 5 min core condition audio segments are reduced to 10 s duration. It is important to note that for TIMIT speaker data is collected in controlled, prompted read, noise-free conditions with a fixed Shure SMIOA head-worn microphone. However, for NIST SRE-2008 data, alternative handsets (landline, cordless, cellphone) as well as microphones are employed. Also, SRE-2008 has multiple languages, as well as language mismatch where a speaker's training data can be in one language, and their test data in a different language. Finally, NIST SRE-2008 data is essentially noise-free, so some form of controlled distortion needs to be introduced. Therefore, these two corpora represent two diverse data sets which can illustrate the potential impact of leveraging speaker/environment spaces for speaker recognition.

Two different noise data sets were used in this study. The first set, referred to as 'diverse' set, consists of five noise types that can be distinctly separated by a human listener. These noise include: (1) a flat communications channel (FLN), (2) a helicopter fly-by (HEL), (3) a large city (LCI), (4) a large crowd (LCR), and (5) the cooling fan of a Sun 4/330 workstation (SUN). These noise types were subjectively scored for their stationarity based on first and second moment analysis on a scale of 1 (wide sense stationary) to 10 (nonstationary) for robust speech recognition using the constrained iterative Auto-LSP enhancement (Hansen and Arslan, 1995). These noise types can also be classified based on their frequency characteristics, as summarized in Table 1.

The second set of noise, (which is referred to as the 'vehicle' set), consists of recordings from inside six different types of vehicles. These include (all models manufactured by General Motors Corporation): (1) Chevy Blazer (BLA) SUV, (2) 4-door passenger Cavalier (CAV) car, (3) an Express (EXP) delivery cargo van, (4) an S10 (S10) compact 2-door pickup truck, (5) a Silverado (SIL) full-size truck, and (6) a Venture (VEN) passenger minivan. Each vehicle noise has eight different sessions. For example, the Silverado has 65 miles per hour (mph) windows-closed, 65 mph windows-open-one-inch, 20–45 mph window-down-one-inch, 20–45 mph windows-closed, 20–45 mph windows-open-halfway, turn signals, idle, accelerate, and these sessions apply to other vehicle noises. Train and test data set are degraded with various noise sessions, and the noise audio never overlaps between train and test data. A more detailed acoustic analysis of the noise for these vehicle types can be found in the study by Hansen (2004). Separate samples of each vehicle noise were used to degrade speakers for train, development, and test, to ensure open segment time variability. All noise files for both noise sets were sampled at 8 kHz and added to the clean TIMIT and SRE 2008 speech at 5 dB SNR.

Table 1
Summary of noise sources. Stationarity is from 1 (stationary) to 10 (non-stationary).

Noise type	Stationarity	Noise group
FLN	1	Broadband
HEL	3	Low frequency band
LCI	3	Time varying colored
LCR	5	Time varying colored
SUN	1	Low frequency

In order to obtain an objective measurement of the difference/separability of the noise types, the Kullback–Leibler (KL) divergence measure was utilized to help quantify the information for discriminating between speaker models. We can estimate the difference between two GMMs using the symmetric KL divergence, which is defined as the sum of the relative entropy between a model pair according to Ben and Bimbot (2003) and Angkititrakul and Hansen (2004):

$$\mathrm{KL}(\Lambda_i, \Lambda_j) = E_{\Lambda_i}\left[\log\frac{\Lambda_i}{\Lambda_j}\right] + E_{\Lambda_j}\left[\log\frac{\Lambda_j}{\Lambda_i}\right], \tag{7}$$

where Λ_i and Λ_j are the speaker model. The individual divergence is computed as Do (2003),

$$E_{\Lambda_i}\left[\log\frac{\Lambda_i}{\Lambda_j}\right] = \frac{1}{2}\left[\log\frac{\det(\mathbf{\Sigma}_i)}{\det(\mathbf{\Sigma}_j)} - \dim(\mathbf{\Sigma}_i) + tr(\mathbf{\Sigma}_j^{-1}\mathbf{\Sigma}_i) + (\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)^t\mathbf{\Sigma}_j^{-1}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)\right]. \tag{8}$$

Table 2 summarizes the resulting log KL divergences for the two sets of noise types used in our experiments. From Table 2, the average "self" log KL divergence is 1.94 (i.e., averaging along the diagonal in bold values), while the average "mismatch" noise distance is 4.12 (i.e., average of all noise-pairs consisting of off-diagonal terms). This table therefore allows us to assess the relative separation of one noise type to another. The individual mismatches within the vehicle set are much closer than that from the diverse noise set.

The key result therefore obtained from the KL divergences is the average distance from a noise type belonging to one set to an alternative noise type, from either set. For matched comparisons, the average log KL divergence is shown to be 1.92 for the diverse noise set and 1.96 for the vehicle noise set. The mismatched average log KL divergence results are summarized in Table 3 (i.e., the diverse–diverse average distortion of −5.31 represents the average of all non-diagonal entries in the upper-left 5 × 5 matrix of Table 2).

These results indicate that while the vehicle noise types are clustered closely together, the diverse noise types are relatively far apart from both the vehicle and other diverse noises themselves. Therefore, one might expect that an algorithm which utilizes information from the background environment of a speaker would have an easier time identifying speakers with noise types from the diverse set, while

Table 2
Summary of KL divergences between noise trained GMMs.

Noise	Diverse set					Vehicle set					
	FLN	HEL	LCI	LCR	SUN	BLA	CAV	EXP	S10	SIL	VEN
FLN	−**1.68**	−5.52	−5.50	−6.96	−5.71	−5.19	−4.89	−4.89	−5.04	−4.69	−5.11
HEL	−4.88	−**2.00**	−4.87	−5.04	−4.82	−4.09	−4.01	−4.02	−4.30	−3.88	−4.29
LCI	−6.00	−5.68	−**1.96**	−5.28	−4.53	−3.48	−3.31	−4.15	−3.80	−3.79	−3.44
LCR	−6.10	−5.43	−4.77	−**2.03**	−5.53	−4.38	−4.48	−4.14	−4.49	−4.31	−4.56
SUN	−7.04	−5.79	−5.05	−6.75	−**1.95**	−4.41	−4.99	−4.59	−4.65	−4.79	−4.70
BLA	−4.80	−4.41	−3.31	−4.45	−3.75	−**1.92**	−2.81	−3.00	−2.83	−3.03	−2.85
CAV	−4.64	−4.16	−3.13	−4.40	−4.02	−2.74	−**1.98**	−2.97	−2.90	−2.93	−2.74
EXP	−4.61	−4.20	−3.85	−4.23	−3.86	−3.00	−3.08	−**1.97**	−2.96	−2.97	−2.85
S10	−4.72	−4.63	−3.56	−4.52	−3.87	−2.81	−2.95	−2.93	−**1.96**	−2.90	−3.05
SIL	−4.36	−3.97	−3.46	−4.19	−3.92	−2.94	−2.93	−2.90	−2.87	−**1.95**	−2.95
VEN	−4.82	−4.51	−3.20	−4.40	−3.82	−2.77	−2.74	−2.78	−3.00	−2.93	−**1.97**

noise types belonging to the vehicle set might produce less potential performance gain since the noise types are closer acoustically. However, there are still distinct differences between the vehicle noises that allow for acoustic model separation.

In this phase, an analysis of vehicle noise effects is performed using a comparison between clean and vehicle noise speech. A total of 20 speakers are randomly selected from the TIMIT corpus, with 10 sentences from each speaker used for analysis. In this scenario, 10 speakers are selected as the target speaker set, and the other 10 speakers become the non-target speaker set. A single vehicle noise is added to each of the target speaker group, and the other five vehicle noises are randomly added to the non-target speaker group (i.e., a single noise type selected from the set of five, and added to a single non-target speaker). Based on the transcribed phone information, the same phonemes are selected from all the speakers. The purpose of this analysis is to compare the inter-speaker differences between clean and vehicle noise speech. To investigate this, we can reduce the potential diversity of the speaker traits by comparing them at the phoneme level. The average log power spectral density (PSD) is computed on each phoneme token, and the PSD of the phoneme from the target and non-target groups is measured by KL divergence. The KL divergence is measured here in the frequency domain since we wish to emphasize the discriminating acoustic speaker space, so the simplified version of the KL from Logan et al. (2001) is used here,

$$\mathrm{KL}(\mathrm{PSD}_i, \mathrm{PSD}_j) = \frac{\boldsymbol{\Sigma}_i}{\boldsymbol{\Sigma}_j} + \frac{\boldsymbol{\Sigma}_j}{\boldsymbol{\Sigma}_i} + (\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)^2 \cdot \left(\frac{1}{\boldsymbol{\Sigma}_i} + \frac{1}{\boldsymbol{\Sigma}_j}\right). \quad (9)$$

The phoneme distance using the KL divergence is a measure of inter-speaker distance between clean and noisy conditions. The results of the divergence measurements are shown in Table 4. The selected phonemes are produced more than 5 times for each speaker so as to capture general speaker variability and reduce phoneme context dependencies. When it comes to KL difference between clean and noisy conditions, the average KL difference of phonemes from the low energy phoneme group is larger than from a high energy phoneme group. This means that the low energy phoneme group provides useful information when comparing clean and noise conditions although this phoneme group has generally been set aside and considered unreliable for traditional clean speaker recognition applications. The largest KL difference occurs in the silence part containing vehicle information.

Table 3
Average mismatch in log KL divergence results.

Noise set	Diverse	Vehicle
Diverse	−5.31	−4.36
Vehicle	−4.13	−2.90

3. Noise and in-set/out-of-set speaker recognition

While most in-set/out-of-set speaker recognition systems (Angkititrakul and Hansen, 2007; Prakash and Hansen, 2007; Angkititrakul and Hansen, 2004; Suh and Hansen, 2012) work reasonably well under clean conditions, the introduction of noise corruption causes a significant change in performance and thus degrades the equal error rate of the system. Numerous techniques have been suggested for general speaker recognition as a means to suppress noise from the speech signal in order to decrease system error; these include spectral subtraction and quantile-based noise reduction (Stahl et al., 2000). Recently, several studies have considered using the noise context as an information source for which the system can adapt its decisions. Akbacak and Hansen (2007) proposed the framework of "Environmental Sniffing" which can detect, classify, and track acoustical environmental structure in order to seek out detailed information that characterizes these conditions and use that knowledge to direct the processing of speech systems. In another example, Müller considered estimating the acoustic context in order to determine whether or not certain acoustic classifiers would be reliable for speaker classification (Müller et al., 2005; Müller, 2007). Müller compare the various classifiers such as Gaussian Mixture Model (GMM), Support Vector

Table 4
Log power spectrum KL divergence measure comparison between clean and noise condition for energy based.

	Phone Class	Phoneme	Number of Phonemes	KL in clean	KL in noisy	KL difference	Avg.	Total avg.
Low energy phonemes	Stops	t	140	277	301	24	27	17
		k	159	275	328	53		
		q	150	281	285	4		
	Fricatives	s	307	309	309	0	7	
		sh	157	308	291	−11		
		f	115	276	315	35		
High energy phonemes	Nasals	m	169	279	308	29	13	15
		n	289	284	281	−3		
	Semivowels glide	l	270	299	284	−15	26	
		r	292	266	286	20		
		w	121	294	366	72		
	Vowels	iy	272	283	291	8	5	
		ih	221	286	293	7		
		eh	154	275	287	12		
		ey	123	283	287	4		
		ae	178	281	280	−1		
		aa	137	280	294	14		
		ah	100	300	315	15		
		ao	113	309	304	−5		
		ax	159	283	286	3		
		ix	361	272	273	1		
		axr	143	285	287	2		
Silence		h#	400	265	350	85	85	85

Machine (SVM), and Neural Network (NN) to extract the context information integrating into speaker recognition.

For the purposes of this study, we assume that the background environment is the same for a particular speaker between train and test. This assumption allows for the use of noise as an aid in the successful acceptance or rejection of an input speaker for in-set/out-of-set speaker recognition. While this assumption cannot be made for every speaker recognition scenario, there are many applications in which it applies. In the case where the rapid detection and tracking of speakers in a relatively short time period is necessary, generally the speaker will be in the same environmental context (i.e., seconds).

There are many applications where rapid detection and tracking of speakers over audio streams is necessary, such as spoken document retrieval, real-time voice dialog, or monitoring pilots during air traffic control. For example, consider the tracking of various TV anchors and correspondents reporting the news. The main anchors will be, with a high degree of certainty, reporting from within the studio, and thus a noise model based on the background acoustics of the studio should be seen both during the training and test phases. Likewise, the traffic correspondent reporting from a helicopter will always have the helicopter as his or her environmental context. Another scenario where this assumption holds is for monitoring and tracking of commercial communications at airports involving ground and air units. Commercial aircraft communications will have pilots in the same aircraft during take-off, taxi (moving on the ground), and landing. The pilot will have a distinctive noise environment when compared to the driver of a baggage transportation vehicle or an air traffic controller. Furthermore, it is highly unlikely that these speakers would switch environments during a restricted time period, since the probability of an individual such as a pilot also being an operator of a ground transport vehicle can safely be assumed to be very small.

It is important to note that the background noise information is not the main focus of an in-set/out-of-set speaker recognition system; but rather, the noise context plays a role since that knowledge can be used to augment the speaker-dependent information the system already employs as a basis for its decision.

4. Leveraged approach: SPKR + ENV

The new approach proposed in this study is to increase performance of in-set/out-of-set speaker recognition by taking advantage of the assumption that a given speaker will remain in the same noise environment between train and test phases. With this scenario as our foundation, we choose to not suppress or ignore the background acoustics, but instead embrace this as potential knowledge which could further improve speaker recognition performance in noisy conditions.

The ‘standard’ input frame selection method used as our baseline is one in which an energy threshold is applied to the speech waveform; frames with an energy above the threshold are used to train a GMM for the enrolled speakers, while frames with energy lower than the threshold are set aside. For clean speech, this baseline system produces an EER of 5.00%. The new method developed here employs an energy-frequency grid to tag input frames. In the simplest case, low energy frames (separated from high energy frames by an energy threshold λ) are used to train a separate GMM which represents the noise or silence content, with some low energy consonant information (see Fig. 2). When an input speech signal is submitted, the system evaluates the log likelihood frame scores associated with both in-set and out-of-set cases for both high energy and low energy GMMs. Next, a weight (β) is applied to the low energy scores and used to combine with those from the high energy GMM to create an overall leveraged (SPKR + ENV) score. The final decision is based on these scores and EER performance is calculated. The system is then enhanced by considering a generalization where the frame scheme grid consists of both energy and frequency.

Speaker recognition systems generally set aside low energy frames, since they normally contain low-energy consonants or silence which are prone to noise. Since our method for in-set/out-of-set speaker recognition considers very small amounts of train (5 s) and test (2 s) data, setting aside any data could lower performance in noise scenarios. Therefore, the proposed method assumes that high energy frames have speaker dependent phonemic content with some background environmental structure, and low energy frames which are expected to have primarily environmental content with some speaker dependent consonant information. We employ an in-set framework similar to our earlier work (Angkititrakul and Hansen, 2007), where the speaker size is 60, with a 15/45 in-set/out-of-set size. By varying the weight value β (see Fig. 2), it is possible to control the emphasis placed on low energy environmental centric versus high energy speaker centric scores.

The results reported in this section are based on a simplified version of the algorithm that does not yet employ frequency analysis/filtering (Section 5 introduces the frequency analysis/filtering phase).

4.1. Evaluation: diverse noise set results

In order to provide a comparative analysis of the benefits of the new (SPKR + ENV) leverage approach, several test sets were randomly created with an equal number of each of the five noise types from the diverse noise set. For the 60 speakers, each of the five noise types were used for a random set of 12 speakers, and the speakers that comprised a certain noise type were kept constant between train and test. The same 15 speakers were chosen as the in-set

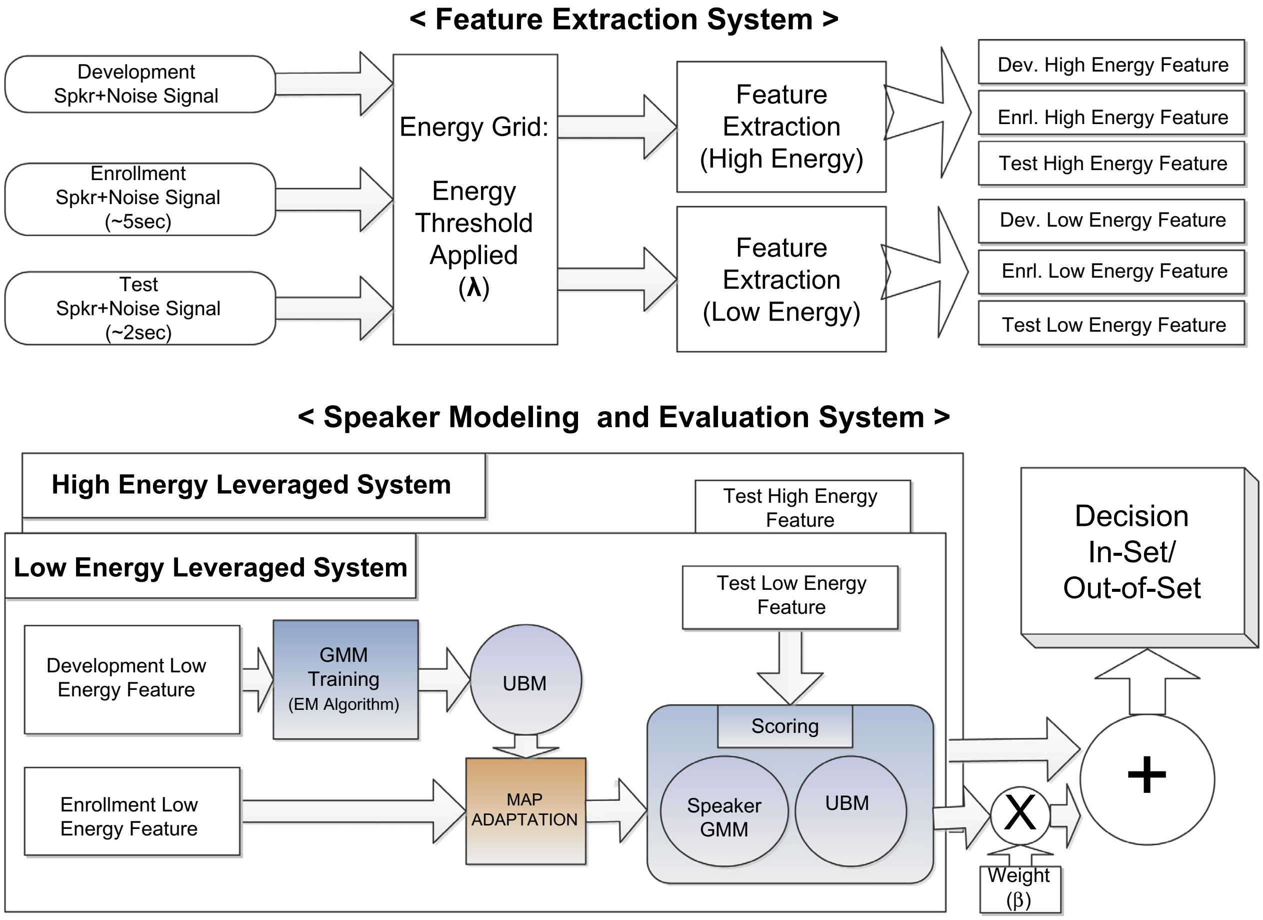
< Feature Extraction System >
Development Spkr+Noise Signal
Enrollment Spkr+Noise Signal (~5sec)
Test Spkr+Noise Signal (~2sec)
Energy Grid:
Energy Threshold Applied (λ)
Feature Extraction (High Energy)
Feature Extraction (Low Energy)
Dev. High Energy Feature
Enrl. High Energy Feature
Test High Energy Feature
Dev. Low Energy Feature
Enrl. Low Energy Feature
Test Low Energy Feature
< Speaker Modeling and Evaluation System >
High Energy Leveraged System
Test High Energy Feature
Low Energy Leveraged System
Test Low Energy Feature
Development Low Energy Feature
GMM Training (EM Algorithm)
UBM
Enrollment Low Energy Feature
MAP ADAPTATION
Scoring
Speaker GMM
UBM
X
Weight (β)
+
Decision In-Set/ Out-of-Set

Fig. 2. Leveraged (SPKR + ENV) system approach for in-set/out-of-set speaker recognition.

speakers for all tests, and eight random noise combinations were used to allow for a more comprehensive evaluation within the in-set scenario.

Once the test sets were constructed, the traditional baseline method which sets aside low energy frames was employed for in-set/out-of-set speaker recognition, yielding an average EER of 9.31% using TIMIT and the 5-type diverse noise set. The baseline method sets the fixed energy threshold with $\lambda = 0.3$ for diverse noise set. Next, the simplest case of the leveraged (SPKR + ENV) approach (no frequency partitioning) was employed, with $\lambda = 0.3$ and with β values ranging from 0.05 to 1.00 (note: λ and β terms are shown in the processing blocks of Fig. 2). It is noted that with $\beta \in (0.05 \leftrightarrow 1.00)$, the log likelihood scores from the GMM trained on low energy frames are counted at most as equal to the scores from the high energy frames ($\beta = 1.0$), with a minimum weight of 5% ($\beta = 0.05$).

For example, if $\beta = 0.50$, and assuming S_1 and S_2 are the high energy and low energy scores, respectively, then the final log likelihood score result is $S_1 + (0.50)S_2$. Thus, the high energy scores comprise $\frac{1}{1.5}$, or 2/3 of the final score, while the low energy score comprises $\frac{0.5}{1.5}$, or 1/3 of the final score. This also allows the researcher/developer to understand from where the most salient/consistent speaker content is located.

The average results from this method (the average of the EERs for each possible β value) results in an EER of 6.78%, with an average relative performance gain of +27.2%. For each of the evaluation sets, the optimum performance occurs when $\beta = 0.65$, with an average EER of 5.28%. These results indicate that for the optimum tested value of β, the absolute improvement (decrease in EER) was 4.03%, with a relative improvement of +43.3%. Additionally, we see that the optimum performance of the SPKR + ENV algorithm approaches that of the clean baseline.

The evaluation process was also performed using an energy threshold of $\lambda = 0.1$. Using this threshold, the resulting average EER is 6.73%, which corresponds to an average relative performance gain of 27.7%. Using the optimum tested value of β, the EER decreases to 5.70%, resulting in an absolute improvement of 3.61%, with a relative improvement of +38.8%. Therefore, while still yielding improved performance over the baseline method, the improvement obtained with an energy threshold of $\lambda = 0.1$ was less than that seen when the frame energy threshold is set to $\lambda = 0.3$. The results of the $\lambda = 0.3$ and $\lambda = 0.1$ experiments are summarized in Table 5, where the average EER rates are recorded. The variance of the EER rates across the tests is also shown, in order to reflect the variability in performance as random noise combinations are used. The results here show significant improvement in system consistency when employing the leveraged SPKR + ENV system.

Fig. 4 shows the detection error tradeoff (DET) curves for the baseline and optimum SPKR + ENV methods for the $\lambda = 0.3$ configuration. This curve demonstrates how the SPKR + ENV algorithm improves overall EER performance for the diverse noise set.

4.2. Evaluation: vehicle noise set results

Similar to the speaker + noise audio test sets created for the diverse noise set, a collection of test sets were also randomly created with the 6 vehicle noise set, where an equal number of noise tracks for each of the noise types were used. For the 60 speakers, a random selection of each of the 6 vehicle noise types were used to acoustically degrade 10 speakers' speech segments for each noise, and those speakers that comprised a certain noise type were kept constant between train and test for both TIMIT and SRE-2008 databases. The same 15 speakers were chosen as the in-set speakers for all tests, and eight random noise combinations were used to increase the experimental file count.

Once the TIMIT test sets were constructed, the traditional baseline method of setting aside low energy frames was employed for in-set/out-of-set speaker recognition, yielding an average EER of 10.14%, which is relatively increased by rate of −8.18% compared the EER of diverse noise set.

Next, the simplest case of the leveraged (SPKR + ENV) approach (no frequency partitioning) was employed, with a frame energy threshold of $\lambda = 0.3$ and with β values ranging from 0.05 to 1.00, where again the log likelihood scores from the GMM trained on low energy frames were counted at most as equal to the scores from the high energy frames ($\beta = 1.0$), with a minimum weight of 5% ($\beta = 0.05$). The average results from this method (the average of the EERs for each possible β value) resulted in an EER of 8.29%, with an average relative performance gain of +18.3%. For each of the evaluation sets, the optimum performance occurred when $\beta = 0.70$, with an average EER of 7.36%. These results indicate that for the optimum tested value of β, the absolute improvement (decrease in EER) was 2.78%, with a relative improvement of +27.4%.

The evaluation process was also performed using an energy threshold of $\lambda = 0.1$. Using this threshold, the resulting average EER is 9.44%, which corresponds to an average relative performance gain of 6.90%. Using the optimum tested value of β, the EER decreases to 7.92%, resulting in an absolute improvement of +2.22%, with a relative improvement of +21.9%. Once again, the improvement obtained with an energy threshold of $\lambda = 0.1$ was less than that for the increased frame energy threshold of $\lambda = 0.3$. The results for the $\lambda = 0.3$ and $\lambda = 0.1$ experiments are summarized in Table 6. Again, there is measurable improvement in overall average EER; however, the reduction in the variance of the EER is not as significant for vehicle (Table 6) versus diverse (Table 5) Noise sets.

Table 5
System comparison for EER and variance for diverse noise set.

	$\lambda = 0.3$		$\lambda = 0.1$	
	Avg. EER	Variance	Avg. EER	Variance
Baseline	9.31	9.85	9.31	9.85
Average SPKR + ENV	6.78	0.51	6.73	0.53
Optimum SPKR + ENV	5.28	0.97	5.7	0.93

Table 6
System comparison for EER and variance for vehicle noise set.

	$\lambda = 0.3$		$\lambda = 0.1$	
	Avg. EER	Variance	Avg. EER	Variance
Baseline	10.14	10.36	10.14	10.36
Average SPKR + ENV	8.29	5.51	9.44	7.21
Optimum SPKR + ENV	7.36	5.96	7.92	7.93

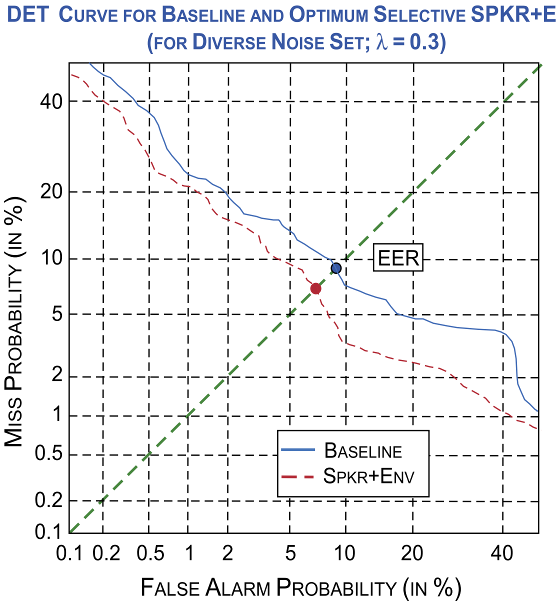

Fig. 3. DET curve for baseline and optimum SPKR + ENV method for the diverse noise set using TIMIT data.

Fig. 3 shows the detection error tradeoff (DET) curves for the baseline and optimum SPKR + ENV methods for the $\lambda = 0.3$ setup. This curve demonstrates how the SPKR + ENV algorithm improves overall EER performance also for the vehicle noise set.

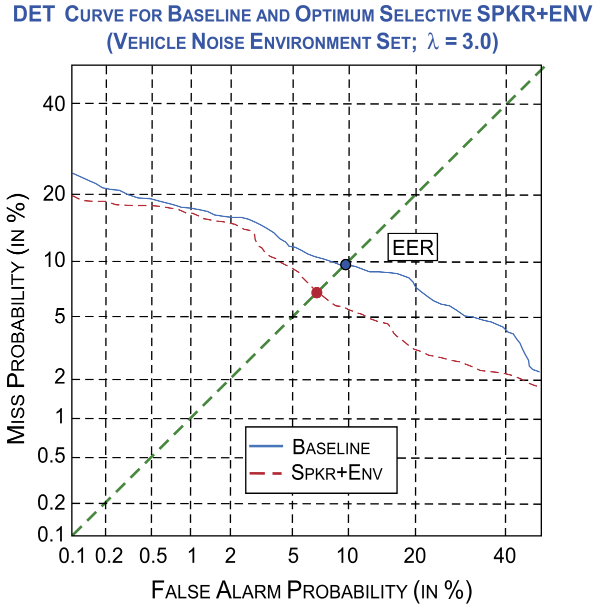

Fig. 4. DET curve for baseline and optimum SPKR + ENV method for the vehicle noise set.

Next, speaker data from the NIST SRE-2008 database were degraded under the same vehicle noise set configuration as that used for TIMIT. Here, the 5 min core set of NIST SRE-2008 database is used to construct a set of 60 speakers, where the duration of data is reduced to 10 s to explore the benefit of the leveraged approach. The frame energy threshold $\lambda = 0.3$ with β values ranging from 0.1 to 0.5 were chosen, and β values over 0.5 does not improve any performance. Table 7 shows the baselines, as well as average and optimum SPKR + ENV leveraged approaches. The experiment is performed to measure the effect of various vehicle noise on recognition. One vehicle noise are added to clean speech for one experiment, and 6 vehicle noise are also added to clean speech to measure the effect of noise on recognition. Silverado vehicle noise is selected for one noise type experiment. The Baseline system only uses the High Energy (HE) part of speech feature, and the prefix word indicates the condition of speech type in experiment in Table 7. Prefix word “Clean” indicates that the system is performed with the only clean speech audio, and the prefix “One” indicates using only one vehicle noise is used for the baseline evaluation. The Baseline without any prefix word uses the 6 vehicle noise types same as TIMIT experiment for vehicle noise. The noise leverage system performs better in both data durations. The absolute +3.89% EER in 10 s and +1.53% EER in 5 min are improved from baseline performance. The one noise and 6 noise types experiment also verifies that the noise also has discrimination ability within limited number of noise types. In the 10 s case, the leveraged approach improves performance gains more than 5 min case since the low energy GMM model helps to provide accurate environmental classification results with the high energy GMM model. It is noted that for the 5 min core case data scenario, the increased amount of data has a clear benefit in more accurately representing both the speaker and environment spaces.

Comparing results from both noise sets, we see that the variance of the SPKR + ENV results for the diverse noise set is an order of magnitude lower than that of the vehicle noise set. This is due to the tighter clustering of the output results, which is believed to be due to noises from the diverse noise set being more distinct from each other.

4.3. Mixture optimization

In order to further refine and optimize the SPKR + ENV framework, the effects of the number of

Table 7
System comparisons of EER for SRE 2008 plus vehicle noise set.

Duration	Clean baseline	One noise baseline	Baseline	Optimum SPKR + ENV
10 s	32.2	48.0	39.7	35.8 ($\beta = 0.5$)
5 min	13.3	33.4	25.4	23.8 ($\beta = 0.1$)

mixtures used in the creation of the In-set/Out-of-set UBM and GMMs were examined. Initially, 32 mixtures were used for the UBM, which was used for modeling both the high-energy speaker dependent frames and the low-energy noise dependent frames. Due to the relatively substantial amount of speech and speaker data available for UBM training, the UBM mixture number was increased to 128. Additionally, what limited speaker content is present in the low-energy noise dependent frames is believed to be insufficient to completely ‘fill’ all 32 mixtures of the GMM. Therefore, the algorithm was evaluated with reduced mixture numbers for low energy (LE) GMM training; specifically, results were obtained for 16 and 8 mixtures. The GMM training for the high energy (HE) speaker dependent frames was kept fixed at 32 mixtures (the four configurations are shown in column 1 of Table 8).

The results from these optimization tests using the vehicle noise set with $\lambda = 0.3$ are summarized in Table 8. These results show that the largest performance increase (decrease in EER) occurs when the UBM mixture order is increased from 32 to 128. An additional slight performance increased is obtained by reducing the mixture order of the low-energy from 32 to 8. Therefore, the optimum results were obtained for a 128 mixture UBM, with a 32 mixture GMM used for high-energy frames and an 8 mixture GMM for low-energy frames during the test phase. The optimum SPKR + ENV EER with optimized mixture order is 6.81%, corresponding to a +5.40% relative improvement over the SPKR + ENV algorithm with an un-optimized mixture order (e.g., case where UBM/HE GMM/LE GMM is 32/32/32).

5. Selective leveraging framework

The next step in enhancing the SPKR + ENV algorithm is to formulate a method in which the noise environments could be evaluated in a manner that allows the decision to strengthen the leveraging process. A framework for this type of selectively leveraged SPKR + ENV system can be developed that evaluates the speakers using a grid with both energy and frequency thresholds, thereby partitioning the dimensions as seen in Fig. 5. The goal of the frequency-energy partitioning is that some partitions will contain more speaker dependent traits while other partitions will contain more noise dependent traits.

An examination was performed for three car noise types (BLA, CAV, and EXP) using frequency partitioning variables of $F1 = 300$ Hz and $F2 = 600$ Hz, and energy partitioning variables of $E1 = 0.1$ (normalized) and $E2 = 0.3$ (normalized). The total number of frames distributed to each of the partitions was tabulated, and the percent of the total frames calculated. Fig. 6 shows the percent of the total frames contained in the Low Frequency, High Energy partition (Cell 1 in Fig. 5) after being passed through a five-point median filter for each utterance. This Fig. 6 clearly demonstrates three distinct bands for the BLA, CAV, and EXP degraded files, confirming that environmental background would be a useful discriminatory trait with frequency dependency. Therefore, it is possible to develop a SPKR + ENV system that performs front-end analysis of the noise content using the grid partitioning method before leveraging the background noise. An application of this analysis is to verify that the speaker stays in the same environment between train and test, and upon verification apply the SPKR + ENV process.

The SPKR + ENV method proposed in Fig. 2 is employed with the frequency-energy partitioning method illustrated in Fig. 5, using the vehicle noise set with partitioning variables of $F1 = 300$ Hz, $F2 = 600$ Hz, $E1 = 0.1$ (normalized), and $E2 = 0.3$ (normalized). The noise dependent frames (Cells 1 and 2 in Fig. 5) were combined, as were the speaker dependent frames (Cells 3 and 4). The weighted scores of the noise dependent frames were then added to those from the speaker dependent frames. This leveraging of the background environment yielded the results summarized in Table 9.

Thus, we see that the noise dependent and mixed partitions (1,2, and 5) provide limited speaker discriminatory performance in terms of in-set speaker identification. The speaker-dependent partitions (3 and 4), however, yield improved results over baseline (+15.8% relative improvement)

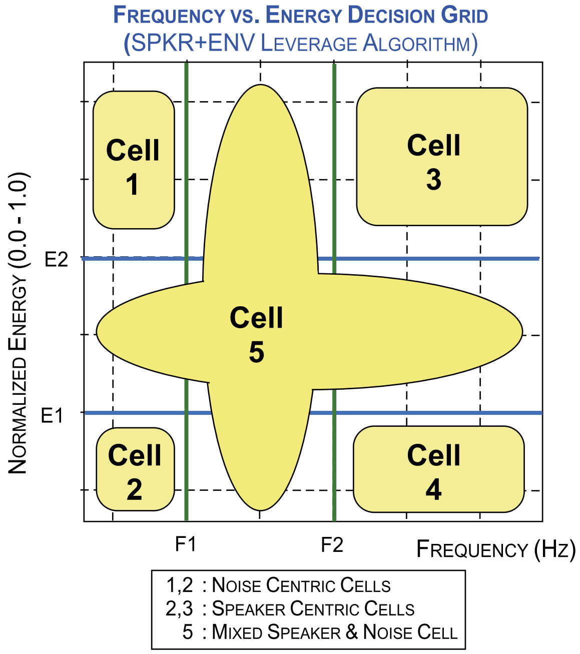
Frequency vs. Energy Decision Grid
(SPKR+ENV Leverage Algorithm)
Normalized Energy (0.0 - 1.0)
E2
E1
Cell 1
Cell 3
Cell 5
Cell 2
Cell 4
F1
F2
Frequency (Hz)
1,2 : Noise Centric Cells
2,3 : Speaker Centric Cells
5 : Mixed Speaker & Noise Cell

Fig. 5. Decision grid consisting of energy versus frequency partitioning.

Table 8
SPKR + ENV mixture optimization EER results for vehicle noise set.

Mixtures (UBM / HE GMM / LE GMM)	Baseline EER	Optimum SPKR + ENV EER	Avg. SPKR + ENV EER
32/32/32	10.14	7.36	8.29
128/32/32	8.75	6.95	7.96
128/32/16	8.75	7.22	8.34
128/32/8	8.75	6.81	7.95

even before leveraging of the background environment. Once the background environment is leveraged by applying the weighted noise dependent frames, the optimum EER drops to 7.22%, which represents a +28.8% relative improvement over the baseline. Fig. 7 shows the resulting detection error tradeoff (DET) curves for baseline and optimum SPKR + ENV methods for the selective frequency-energy leveraging framework. This curve demonstrates how the selective SPKR + ENV algorithm improves performance across a range of operating points.

Unfortunately, the average SPKR + ENV results (the average EER across all possible values of β) are actually worse than the standard baseline. This is because the selective leveraging approach utilizing the frequency and energy partitioning grid requires a much more fine tuned optimization when compared to the simplified version that does not account for frequency. This issue is illustrated clearly in Fig. 8.

Fig. 8 shows that while the EER result of each β value for the simplified approach is relatively independent of the value of β, the same cannot be said for the selective approach; rather, the EER of the selective approach dramatically increases as β is increased. Therefore, it is much more important that the selective approach framework be optimized (which in its most basic form means utilizing only small values of β).

Furthermore, the selective leveraging approach was applied with only Cell 1 (see Fig. 5) used as the noise dependent frames which were weighted and combined with the speaker dependent frames. This was done in an attempt to determine if the low energy, low frequency frames in Cell 2 (see Fig. 5) were to blame for the extreme dependency of the EER on the value of β. This experiment results in an optimum average EER of 6.94%, corresponding to a +31.6% improvement over the baseline (an additional 2.8% over the approach utilizing both Cells 1 and 2). However, the results were just as dependent on the value of β, indicating that the selective leveraging approach still requires a refined optimization process. Fig. 9 shows the final DET curves for the standard baseline and optimum SPKR + ENV methods for this version of the selective leveraging framework.

It is important to note that while at first glance the average results across all possible values of β are not effective, merely restricting the range of β can significantly improve performance. Specifically, by restricting β to be in the range of 0.05–0.20 (as opposed to the normal range of 0.05–1.00), the average SPKR + ENV EER drops to 9.10% when Cells 1 and 2 are used, and 8.75% when only Cell 1 is used.

Table 9
Selective leveraging results.

Evaluation setting	Overall average EER(%)
Original baseline	10.14
Using only Cells 1 and 2	42.00
Using only Cells 3 and 4	8.54
Using only group 5	43.42
Average SPKR + ENV w/ freq. – energy part.	15.90
Optimum SPKR + ENV w/ freq. – energy part.	7.22

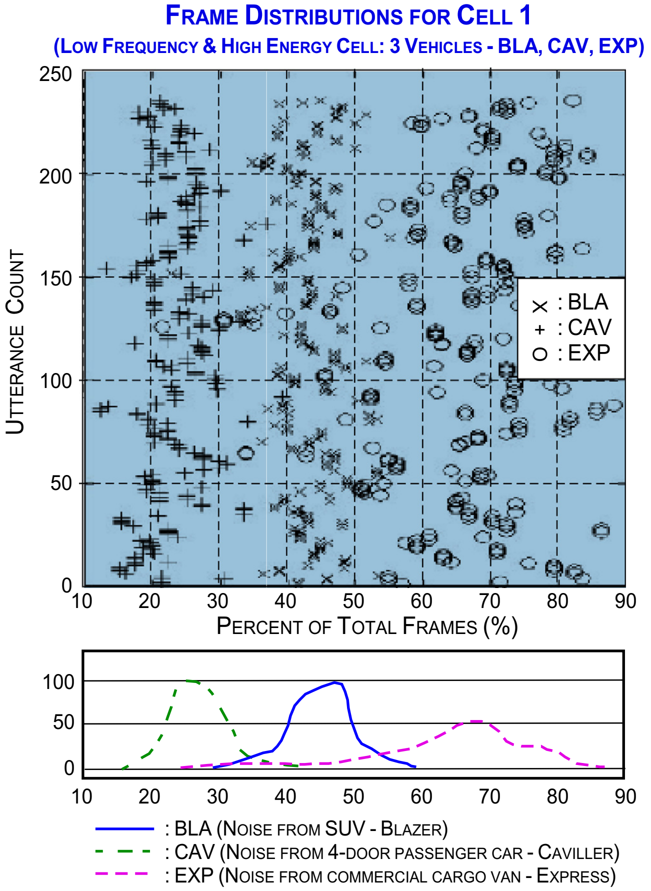

Fig. 6. Frame distribution for low frequency/high energy partition (Cell 1) for BLA, CAV, and EXP files.

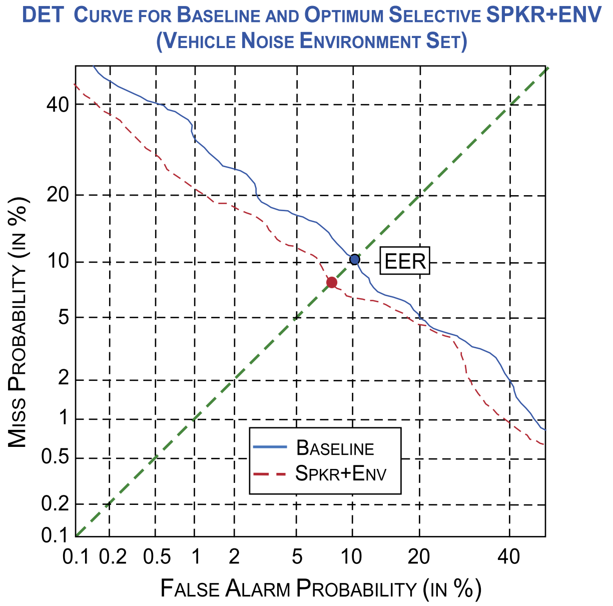

Fig. 7. DET curve for baseline and optimum selective SPKR + ENV method for vehicle noise set.

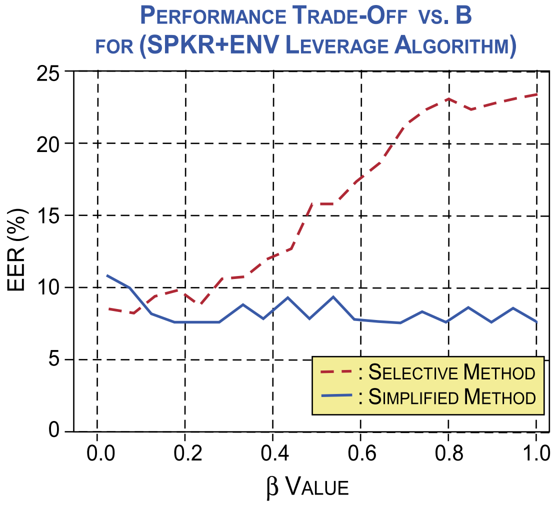

Fig. 8. Optimization comparison of selective and simplified approaches.

Therefore, by restricting the possible values of β, we eliminate a significant portion of the problems that arise from the full selective approach being more sensitive than the simplified approach. We again emphasize the fact that all in-set speaker models were trained with only 5 s of data, and tests performed using 2 s duration files.

6. Discussion and conclusions

Speaker recognition systems must overcome a range of issues in order to achieve and maintain performance in diverse environments. For in-set/out-of-set speaker recognition, where limited amounts of train/test data are generally available, leveraging knowledge of the acoustic environment offers an additional dimension to improve overall system performance. The proposed SPKR + ENV approach shows significant performance improvement by leveraging environmental structure, which virtually all other algorithms intentionally ignore.

By taking into account low energy noise-dependent frames, we significantly increase the performance for in-set/out-of-set speaker recognition. For a simplified version of the algorithm that did not partition the frequency domain, the optimum relative performance improvement was +43.3% for the diverse noise set comprised of noticeably different noise types, and +27.4% on the vehicle noise set comprised of six different vehicle noises (further improved to +32.8% by optimizing the GMM mixture order, where speaker data was drawn from TIMIT) . The performance using 10 s from NIST SRE-2008 shows a +10.8% relative improvement in EER over baseline. For the full selective leveraging approach, the optimum relative performance improvement was +28.8% (low energy, low frequency frames from Cell 2 leveraged) or +31.6% (low energy, low frequency frames ignored).

These results indicate that for situations in which one can assume that the background environment for a speaker remains constant (but randomly distributed across the in-set speakers) between train and test phases, the leveraged (SPKR + ENV) approach will eliminate roughly between (1 out of every 4) to (1 out of every 3) decision errors. Since we are focused on a scenario in which both train and test data are of short duration, this reduction in error is particularly beneficial because of the limitation a lack of data places on other methods such as spectral subtraction.

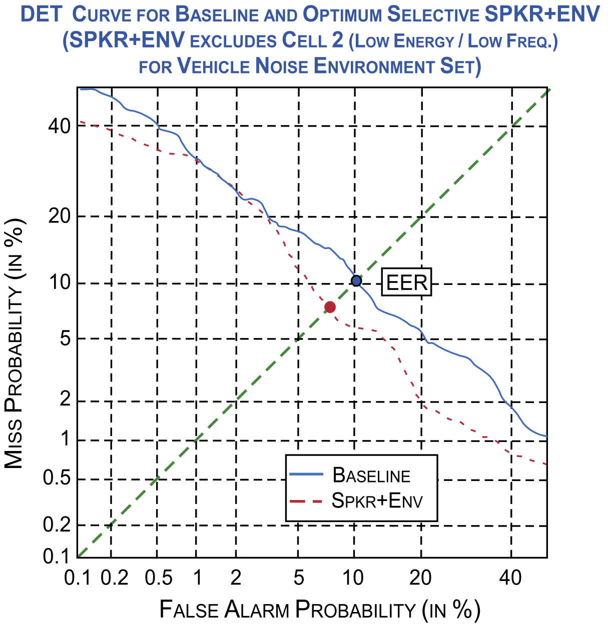

Fig. 9. DET curve for baseline and optimum selective SPKR + ENV method excluding low energy/low frequency content (Cell 2) for vehicle noise set.

One important aspect of error reduction is whether the errors fall into either false accept (FA) or false reject (FR) categories. If the models for two speakers are similar, but their acoustic backgrounds are different, knowledge of the background noise should help drive the models farther apart. An analysis of errors from both the baseline and the SPKR + ENV algorithm shows that the leveraged method eliminates many of the false accept errors.

Several further enhancements are possible given the framework of the SPKR + ENV algorithm. In particular, optimization of the frequency partitioning threshold, and the formulation of an algorithmic approach to adaptively optimize the frequency and energy thresholds for an unknown background environment are options. It was also demonstrated that the selective leveraging approach requires a finer tuned optimization process in order to achieve comparable results to the simplified approach; however, merely restricting the possible range of β significantly improves performance and removes much of the optimization requirement.

One of the major focuses for future work on the SPKR + ENV approach is to determine the optimized energy and frequency thresholds for a wide variety of noise sources. Once this task is completed, if a speaker with a “new” arbitrary noise environment is present in the test phase, the SPKR + ENV algorithm can attempt to classify the acoustic environment present by comparing it to environments that have previously been optimized. Once the closest a priori environment is determined, the optimized thresholds are applied to the new acoustic background. This enhancement would dramatically increase the robustness of the SPKR + ENV algorithm and allow its performance improvements to be experienced over a much wider range of possible noise contexts.

While the leveraged SPKR + ENV approach presented here cannot be applied to every in-set/out-of-set speaker recognition scenario, it does significantly improve performance when utilized in systems focused on the rapid detection and tracking of speakers that remain in the same noise environment between train and test phases. It also allows us to achieve an upper bound on performance improvement when noise is present.

References

Angkititrakul, P., Hansen, J.H.L., 2007. Discriminative in-set/out-of-set speaker recognition. IEEE Trans. Audio Speech Lang. Process. 15 (2), 498–508.

Prakash, V., Hansen, J.H.L., 2007. In-set/out-of-set speaker recognition under sparse enrollment. IEEE Trans. Audio Speech Lang. Process. 15 (7), 2044–2052.

Suh, J-.W., Hansen, J.H.L., 2012. Acoustic hole filling for sparse enrollment data using a cohort universal corpus for speaker recognition. J. Acoust. Soc. Am. 131 (2), 1515–1528.

NIST SRE: U.S. National Institute of Standards and Technology. Speaker Recognition Evaluation, Sep. 3, 2009 (Nov. 14, 2011). <http://www.itl.nist.gov/iad/mig/tests/sre/>.

Shao, Y., Srinivasan, S., Wang, D., 2007. Incorporating auditory feature uncertainties in robust speaker identification. In: IEEE ICASSP 2007, vol. 4, pp. 277–280.

Rose, R., Hofstetter, E., Reynolds, D., 1994. Integrated models of signal and background with application to speaker identification in noise. IEEE Trans. Speech Audio Process. 2 (2), 245–257.

Ariyaeeinia, A., Fortuna, J., Sivakumaran, P., Malegaonkar, A., 2006. Verification effectiveness in open-set speaker identification. IEEE Proc. Vision Image Signal Process. 153 (5), 618–624.

Auckenthaler, R., Carey, M., Lloyd-Thomas, H., 2000. Score normalization for text-independent speaker verification systems. Digital Signal Process. 10 (1-3), 42–54.

Doddington, G., 1985. Speaker recognition identifying people by their voices. Proc. IEEE 73 (11), 1651–1664.

Reynolds, D., Quatieri, T., Dunn, R., 2000. Speaker verification using adapted Gaussian mixture models. Digital Signal Process. 10 (1-3), 19–41.

Xiang, B., Berger, T., 2003. Efficient text-independent speaker verification with structural gaussian mixture models and neural network. IEEE Trans. Speech Audio Process. 11 (5), 447.

Garofolo, J.S., 1993. TIMIT Acoustic–phonetic continuous speech corpus. Linguistic Data Consortium.

NIST SRE: U.S. National Institute of Standards and Technology. The NIST Year 2008 Speaker Recognition Evaluation Plan, Nov. 4, 2008 (Nov. 14, 2011). <http://www.itl.nist.gov/iad/mig/tests/sre/2008/index.html>.

Hansen, J.H.L., Arslan, L., 1995. Robust feature-estimation and objective quality assessment for noisy speech recognition using the credit card corpus (CCDATA). IEEE Trans. Speech Audio Process. 3 (3), 169–184.

Hansen, J.H.L., Huang, R., Zhou, B., Seadle, M., Deller, J., Gurijala, A., Kurimo, M., Angkititrakul, P., 2005. Speechfind: Advances in spoken document retrieval for a national gallery of the spoken word. IEEE Trans. Speech Audio Process. 13, 712–730.

Hansen, J.H.L., 2004. Getting started with the CU-move corpus. Robust Speech Processing Group-CSLR, Univ. of Colorado.

Ben, M., Bimbot, F., 2003. D-MAP: A distance-normalized MAP estimation of speaker models for automatic speaker verification. In: IEEE ICASSP 2003, vol. 2, pp. 69–72.

Angkititrakul, P., Hansen, J.H.L., 2004. Identifying in-set and out-of-set speakers using neighborhood information. In: IEEE ICASSP 2004, vol. 1, pp. 169–184.

Do, M., 2003. Fast approximation of Kullback–Leibler distance for dependence trees and hidden Markov models. Signal Process. Lett. IEEE 10 (4), 115–118.

Logan, B., Salomon, A., 2001. A music similarity function based on signal analysis. In: IEEE Internat. Conf. on Multimedia and Expo, vol. 0, p. 190.

Stahl, V., Fischer, A., Bippus, R., 2000. Quantile based noise estimation for spectral subtraction and Wiener filtering. In: IEEE ICASSP 2000, vol. 3, pp. 1875–1878.

Akbacak, M., Hansen, J.H.L., 2007. Environmental sniffing: Noise knowledge estimation for robust speech systems. IEEE Trans. Audio Speech Lang. Process. 15 (2), 465–477.

Müller, C., 2005. Estimating the Acoustic Context to Improve Speaker Classification. German Research Center for Artificial Intelligence.

Müller, C., 2007. Speaker Classification: Fundamentals, Features, and Methods. Springer, vol. 1, Ch. 2.