Odyssey 2014:
The Speaker and Language Recognition Workshop
16-19 June 2014, Joensuu, Finland

ROBUST LANGUAGE RECOGNITION BASED ON DIVERSE FEATURES

*Qian Zhang, Gang Liu, John H.L. Hansen**

Center for Robust Speech Systems (CRSS), Erik Jonsson School of Engineering,
University of Texas at Dallas, Richardson, Texas, U.S.A.
{qian.zhang, gang.liu, john.hansen}@utdallas.edu

10.21437/Odyssey.2014-24

ABSTRACT

In real scenarios, robust language identification (LID) is usually hindered by factors such as background noise, channel, and speech duration mismatches. To address these issues, this study focuses on the advancements of diverse acoustic features, back-ends, and their influence on LID system fusion. There is little research about the selection of complementary features for a multiple system fusion in LID. A set of distinct features are considered, which can be grouped into three categories: classical features, innovative features, and extensional features. In addition, both front-end concatenation and back-end fusion are considered. The results suggest that no single feature type is universally vital across all LID tasks and that a fusion of a diverse set is needed to ensure sustained LID performance in challenging scenarios. Moreover, the back-end fusion also consistently enhances the system performance significantly. More specifically, the proposed hybrid fusion method improves system performance by +38.5% and +46.2% on the DARPA RATS and the NIST LRE09 data sets, respectively.

Index Terms— Language identification, acoustic feature, fusion, RATS, NIST LRE

1. INTRODUCTION

Language identification (LID) has recently emerged to be of substantial interest in the speech processing community [1-5]. It is a necessary pre-processing component for most automatic speech recognition (ASR) tasks. In recent years, acoustic and phonotactic models have been widely used for LID with some success. Phonotactic approaches usually are based on various phone recognizers and phoneme n-gram statistics to extract discriminative information related to each particular language. The most popular phonotactic modeling techniques are Parallel Phone Recognition and Language Modeling (PPRLM) [6] and Phone recognition-SVM [7]. However, phonotactic models usually only perform well on relatively clean speech. In contrast, acoustic systems are usually based on some spectral features, which are followed by effective analysis models, such as joint factor analysis (JFA) [8] and i-Vectors [9, 10] that can extract valid information efficiently. i-Vector, which has become a popular technique used for different verification and recognition tasks [11-15], can represent each conversation in parallel with a set of low-dimensional total variability factors and demonstrates session variation robustness. Therefore, i-Vector is the analysis model adopted in this study.

In recent studies, Mel-frequency cepstral coefficients (MFCC)-based features were widely employed for LID [1, 2, 5]. However, there are a variety of acoustic features other than MFCC [16-22], which have been successfully utilized for other audio based identification tasks [23-25] but seldom explored for LID. Even on the large-scale task LRE (language recognition evaluation), only several classical features have been involved [26]. While among those distinct features, Some feature sets may perform well in relatively clean conditions but are seldom validated under more challenging conditions (for example, a noisy scenario), which also warrants further investigation here.

In terms of back-end classifiers, various identifiers dependent on front-end feature extraction algorithms have been explored, including artificial neural networks, Gaussian mixture models, support vector machines, etc. Due to the effective performance for i-Vector based systems, generative Gaussian back-ends [27] are used as a benchmark classifier. As a comparison, another newly proposed back-end is examined to advance the investigation.

The main purpose of this study is to systematically investigate different front-end performances on noisy or highly channel-mismatched data and large-scale data, which has not been well investigated. Their performances are validated on two back-ends. The merit of both front-ends and back-ends is further leveraged.

This paper is organized as follows: Sec. 2 explores the features and their configurations, and Sec. 3 describes the principles of the system back-ends. The fusion scheme is detailed in Sec. 4. In Sec. 5, we illustrate the special properties of the corpora, and the comprehensive experimental set-up is shown in Sec. 6. Sec. 7 analyzes the results, and Sec. 8 summarizes the findings.

2. FEATURES

Historically, feature research for LID has focused on identifying a single, universally successful feature. There has been little effort to leverage multiple features across fused LID systems. A detailed investigation is warranted. To be specific, three types of features are explored in our LID experiments.

- **Classical features**
  Mel-frequency cepstral coefficients (MFCC)
  Perceptual linear predictive (PLP)
  Linear frequency cepstral coefficients (LFCC)
  Gammatone frequency cepstral coefficients (GFCC)
- **Innovative features**
  Power Normalized Cepstral Coefficients (PNCC)
  Perceptual minimum variance distortionless response (PMVDR)

* This project was supported by AFRL under contract FA8750-12-1-0188 and partially by the University of Texas at Dallas from the Distinguished University Chair in Telecommunications Engineering held by J.H.L. Hansen.

Table 1. *Configuration of various LID features ('Default' means the original setting for the feature by their authors).*

Feature	Special Configuration	Coefficients(Dim)
MFCC	# of Channels = 26	12
PLP	# of Channels = 20	13
LFCC	# of filter banks = 32	20
GFCC	Default	12
PNCC	Default	12
PMVDR	Default	12
RASTA-PLP	Default	9
RASTA-LFCC	Default	20
Multi-peak MFCC	K=8	13
Thomson MFCC	K=8	13
SWCE_MFCC	K=8	13

- **Extensional features**
  RASTA-PLP
  RASTA-LFCC
  Multi-peak MFCC
  Thomson MFCC
  Sine-weighted cepstrum estimator (SWCE) MFCC

The MFCCs and PLP coefficients are well-known acoustic features that are used widely for many speech processing tasks. The LFCC, which can be extracted using a linear filter bank instead of a Mel scale filter bank, has been shown to be superior for speaker recognition [19]. LFCCs work especially well for female speakers because their shorter vocal tract results in the higher formant frequency, which can then be easily captured by a linear filter bank. For LID task, the overall performance can be boosted by improving the performance on female speech sub-set. Similar to the MFCCs, the Gammatone feature (GF) is obtained from a bank of Gammatone filters, which were originally proposed to model human cochlear filtering. The GFCCs are derived from the GF by a discrete cosine transform (DCT). It has been proved that the GFCC could provide a substantial contribution to the noise robustness of the Speaker Identification (SID) system in [17], but this contribution has yet to be validated in LID.

In addition to those classical features, researchers have recently also proposed some innovative features that could address various types of background noise or room reverberance. Instead of using the traditional log nonlinearity as in the MFCC, the PNCC [21] use a power-law nonlinearity in a medium-time power analysis that combines traditional short-time processing with a noise-suppression algorithm, which is realized by an asymmetric filter that suppresses background excitation. Similarly, the PMVDR [16] cepstral coefficients were proposed by obtaining a minimum variance distortionless response spectral estimator, which represents the upper envelope of the speech signal. Unlike the MFCC, which utilizes a filter-bank, the PMVDR performs warping on the FFT power spectrum directly. This method is desirable for speaker independent tasks (such as a LID) because less pitch information is included. The advantage of this approach is prominent, especially in noisy car environments, gender mismatches, or imbalances in the training data.

In addition to exploring new features, researchers have also considered using different filters that have the potential to improve classical features by offering a degree of noise suppression. The RASTA filter is a special band-pass filter, which suppresses spectrally high or low derivative (i.e., very rapidly-changing or very slow-changing) components versus a typical spectral range of speech. We used the RASTA-PLP and the RASTA-LFCC combinations for our system. In addition, we employed a multi-taper method, which applied multiple (K) uncorrelated windows (tapers) to process a signal in the time-domain, and then averaged the signal in the frequency domain to more accurately estimate the power spectrum of the signal. For example, we represented the short-term signal spectrum using MFCCs, which were computed from a windowed discrete Fourier transform (DFT). Although windowing reduces spectral leakage, the spectrum estimation variance remains high; therefore, extensional features utilizing a multi-taper method is proposed to solve this problem. Except for the Hamming window, there are several alternative filters that have demonstrated a benefit for SID, such as multi-peak, Thomson, and the SWCE. All of these filters can extract the short duration features in a manner similar to the MFCCs, but with a much lower variance. For the Shifted Delta Cepstra (SDC) [18], previous work has shown that incorporating additional temporal information will benefit an acoustic event identification system, such as the systems of emotion identification [24-25], SID, and LID. Therefore, the SDC based on the common scheme, [7-1-3-7], is applied to all of the above-mentioned feature extraction, to derive 56 dimensional raw features used for i-Vector extraction (see Sec. 6).

Furthermore, because distinct feature sets have their own characteristics and advantages that might complement each other, feature level fusion will also be explored. The configuration employed in this study is shown in Table 1. Again, we note that this study is not focused on finding a single more effective feature for LID but on exploring a systematic strategy to leverage multiple features that are complementary. It is believed that this strategy will help to reduce the impact of recording mismatch conditions in both the training and the test sets to improve the LID performance.

3. BACK-END

Two back-end classifiers are investigated here. The first is the generative Gaussian back-end, which is the classical classifier for i-Vector based LID system. The second is the Gaussianized cosine distance scoring (GCDS) method, which was recently proposed to address multiple enrollment session-based tasks in [9].

3.1. Gaussian Back-end (GB)

For the Gaussian back-end, the distribution of i-Vectors for each language was modeled by a Gaussian distribution, where a full covariance matrix was shared across all of the languages. For each i-Vector ω corresponding to a test utterance, we evaluated the log-likelihood for each language as:

$$\ln p(\omega \mid l) = -\frac{1}{2}\omega^T \sum{}^{-1}\omega + \omega^T \sum{}^{-1} m_l - \frac{1}{2} m_l{}^T \sum{}^{-1} m_l + c \qquad (1)$$

where m_l is the mean vector for language l, Σ is the common covariance matrix, and c is a constant. To enhance the efficiency of the operation and to suppress redundant information, a dimension reduction based on linear discriminative analysis (LDA) was also applied before the GB. The maximum number of dimensions for identification was one less than the number of classes.

3.2. Gaussianized cosine distance scoring (GCDS)

It is noted that the performance of the classical within the class covariance normalization (WCCN) based Cosine Distance Scoring (CDS) method depends strongly on the WCCN projection, which is usually difficult to estimate (especially in noisy and/or channel mismatch conditions). Therefore, we recently proposed to replace

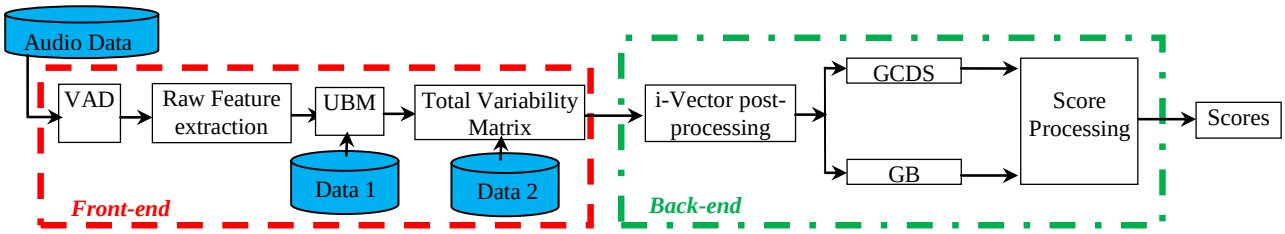

Figure 1: *i-Vector based Language identification system block diagram. Data 1 and 2 correspond to raw feature data for the UBM and Total Variability matrix, respectively. In this study, Data 1 and Data 2 are the same group of data as the training data. 'Audio data' is all of the acoustic data involved for the verification task.GB represents the Gaussian Back-end, GCDS is the Gaussianized CDS.*

the WCCN with background data-based Gaussianization, called Gaussianized CDS (GCDS[1]) [9].

4. FUSION

System fusion usually significantly benefits the overall system performance because of the complementary effect among each individual session [28, 29]. Therefore, we propose to leverage the impact of both the front-end and the back-end fusion. The front-end fusion is implemented by concatenating the i-Vectors from different raw features (also known as feature concatenation). The back-end fusion is implemented by utilizing the FoCal multi-class toolkit [30]. By measuring the goodness of each recognizer and assigning a proper weight based on the supervised development data, FoCal (linear logistic regression fusion) provides a calibrated fusion of the scores of multiple recognizers. In contrast to binary logistic regression, multi-class logistic regression scheme requests regularization, which involves parameter tuning. The parameters lambda and epsilon are set as 0.4 and 0.1, respectively.

5. CORPORA

To explore the system capability in identifying the language under highly degraded and/or noisy communication channel conditions, we performed an evaluation on the DARPA-sponsored Robust Automatic Transcription of Speech (RATS) database[2]. The task was to distinguish and identify six highly confusing language categories: (1) Arabic, (2) Farsi, (3) Dari, (4) Pashto, (5) Urdu, and (6) 10 other non-target languages. At the same time, there were severe channel mismatches among all instances, as shown in Fig. 2. Because the testing files were sorted in the same order as the class IDs (Y axis in Fig. 2), assuming the channel has no impact on the experiments, the main diagonal line should be the only red zone with high probability, which indicates close match purely resulted from language factors. However, the confusion analysis shows that the transmission channel factors significantly impacted the classification results in a negative way.

Alternatively, to further validate the performance of the abovementioned features, we also evaluated them on the NIST LRE09 corpus, which contains 23 different languages (only in-set languages are considered here). While there was more language variety, another common challenge was the speech duration mismatch. Data from the training set only consisted of 30 sec. utterances, while the test set were made up of 3, 10 and 30 sec. utterances. Most of the training data were extracted from the VOA (Voice Over America) broadcasts [31]. There were channel mismatches across the training and the test set. The related corpus statistics information is summarized in Table 2, from which we can see that RATS was very noisy, while LRE09 was relatively clean.

Table 2. Corpus statistics for the DARPA RATS and NIST LRE09.

Corpus	DARPA RATS		NIST LRE09(Inset)	
Data set	TRAIN	TEST	TRAIN	TEST
Count	12035	877	11158	31178
Avg. Duration(*sec./file*)	58.3	18.0	39.3	12.4
SNR(dB)	5.9	8.0	23.30	23.8

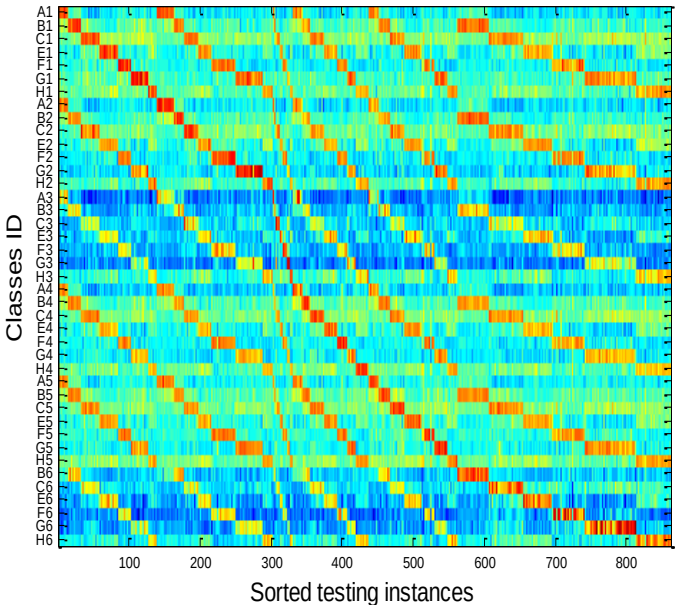

Figure 2. *Channel confusion analysis for RATS corpus (as for class ID, the first character indicates channel label, the second presents 6 distinct language classes. Testing instances are sorted by Class ID).*

6. LID EXPERIMENTAL SET-UP

The modeling was based on an i-Vector framework. The system employed in this study is illustrated in Fig. 1. We performed voice activity detection on all of the audio files with [32]. Specific raw

[1] "GCDS Algorithm source code". [Online]. Available: www.utdallas.edu/~gang.liu/code.htm

[2] Given seven background channels (A, B, C, E, F, G, H) and six language classes, forty-two different combinations can be explored. However, because none of the non-target language examples from the test data were recorded under channel A, there were forty-one categories in total to investigate.

Table 3. *Performance on RATS and LRE09 database (C_{avg}*100).*

Feature category	Feature type	RATS			LRE09		
		GB	GCDS	Backend fusion	GB	GCDS	Backend Fusion
Classical features	MFCC	15.6	14.9	12.8	15.8	14.0	11.5
	LFCC	16.5	16.0	15.0	16.6	15.1	12.2
	PLP	18.7	17.6	16.5	18.7	16.0	13.9
	GFCC	14.9	14.5	13.1	16.6	14.8	12.0
Innovative features	PNCC	14.4	14.0	11.6	16.0	15.1	12.2
	PMVDR	19.1	19.0	16.7	17.8	15.8	13.5
Extensional features	RASTA-LFCC	15.2	13.7	11.3	16.7	14.1	11.3
	RASTA-PLP	14.1	14.2	12.7	15.6	13.3	10.5
	Multi-peak MFCC	15.5	14.0	12.8	15.5	13.7	10.8
	Thomson MFCC	16.0	13.7	12.1	15.7	13.7	11.1
	SWCE MFCC	15.3	13.5	11.5	15.5	13.7	10.9
Feature concatenation		13.1	12.4	9.6	11.7	11.3	8.5

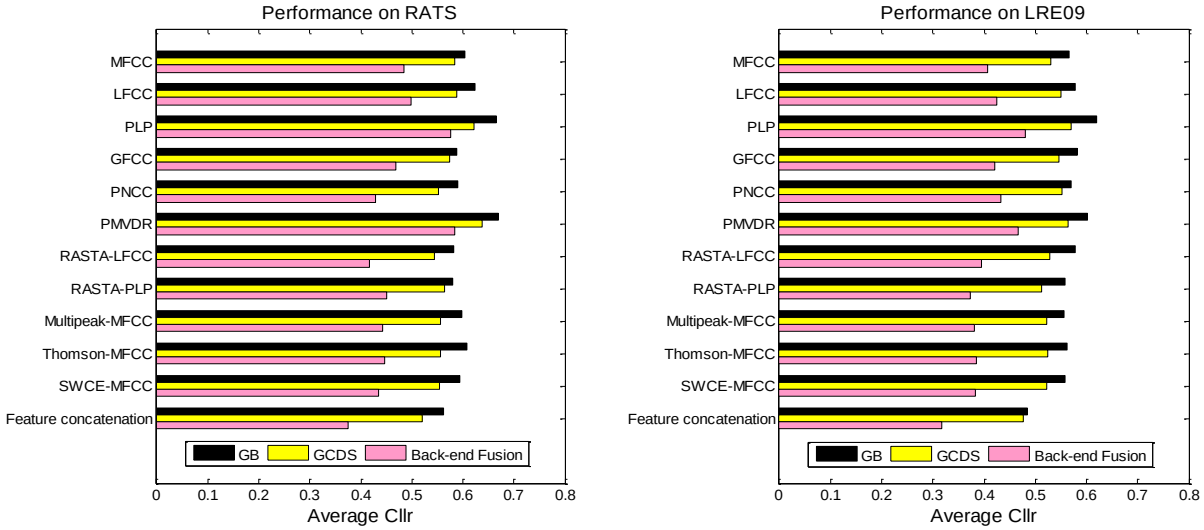

Figure 3. *System performance comparison across two corpora (C_{llravg}).*

features were extracted as described in Sec. 2. After that, the training data from all target languages were used to build a 1024-mixture Universal Background Model (UBM). We followed the i-Vector system paradigm for language recognition as presented in [33, 34]. In the implementation, the total variability matrix was trained on all of the development data (same data as for the UBM) using the EM algorithm for Eigenvoice presented in [35]. We applied 5 iterations for the estimation and then 400 dimensional i-Vectors were derived for further processing [36]. Length normalization was employed in i-Vector post-processing [37].

7. RESULTS

This section presents the experimental verification results for both the individual system and the fused system across two different corpora. In addition, the system performances are evaluated according to two measurements, C_{avg} and C_{llravg}, which are defined in the 2009 NIST Language recognition evaluation (LRE) plan [38]. To make it clear and concise, we only show the average cost performance C_{avg} expression here

$$C_{avg} = \frac{1}{N_L}\sum_{L_T}\left\{\begin{array}{c} C_{\text{Miss}}\cdot P_{\text{Target}}\cdot P_{Miss}(L_T) \\ +\sum_{L_N} C_{\text{FA}}\cdot P_{\text{Non-target}}\cdot P_{FA}(L_T,L_N) \\ +C_{\text{FA}}\cdot P_{\text{Out-of-set}}\cdot P_{FA}(L_T,L_O) \end{array}\right\} \quad (2)$$

where N_L is number of languages in the (closed-set) test, L_O is the Out-of-set language, $P_{Non\text{-}target}$ and $P_{Out\text{-}of\text{-}set}$ are defined as below.

$$P_{\text{Non-target}} = (1 - P_{\text{Target}} - P_{\text{Out-of-set}})/(N_L - 1) \quad (3)$$

$$P_{\text{Out-of-set}} = \begin{cases} 0.0 & \text{for the closed - set condition} \\ 0.2 & \text{for the open - set condition} \end{cases} \quad (4)$$

Here, we only consider the closed-set tasks. In addition to the classification accuracy, the log likelihood ratios based performance

Table 4. *Performance for different test duration condition on LRE09 database. (Hybrid fusion is the backend-fusion based on feature concatenation. "Mixed" means all testing files are considered as a whole and no distinction among test durations).*

C_{avg}*100	3sec	10sec	30sec	Mixed
Baseline(MFCC+GB)	23.1	15.1	9.3	15.8
Hybrid Fusion	13.9	7.2	4.5	8.5
Relative Gain (%)	39.8	52.3	51.6	46.2

measure C_{llravg} (average Cllr) could show more information about system calibration. To illustrate the merit of both feature concatenation and back-end fusion precisely and respectively, we investigated the system performance on four different levels.

- Individual system based on each feature
- Backend-fusion based on each feature
- System based on Feature concatenation
- Backend-fusion based on Feature concatenation (Hybrid fusion)

All of the results are represented as C_{avg} in Table 3. And to make it easier to observe and compare, we show all the performance as C_{llravg} (which has similar trend as C_{avg}) in Fig. 3. Here, i-Vector system that based on MFCC features with GB is used as benchmark system for further performance comparison.

Considering the results based on two different measurements, it is interesting to notify the performance similarities across diverse corpora. Firstly, the GCDS back-end consistently outperforms GB in general (only one exception is applying feature RASTA-PLP for RATS based on measurement C_{avg}). Secondly, either adopting feature concatenation or backend-fusion, system performance will be benefited significantly. Thirdly, when frond-end and back-end fusion are combined together (also known as proposed hybrid fusion), the system performance get improved further and achieve the best results.

In terms of feature exploration, we note that RASTA filter solution has a positive impact on the LFCC and PLP features. To be specific, RASTA-LFCC and RASTA-PLP are the best features for backend-fusion systems according to RATS and LRE09 corpus, respectively. While for front-end fusion, we observe that applying a LDA before concatenation, which reduces the feature dimension in advance, will benefit the system for the LRE09 corpus. However, for the noisy RATS corpus, keeping the feature with the original, higher-dimension achieves a significant improvement. The assumption here is that for the features extracted from noisy data, the discriminative information spreads out across all feature dimensions and is therefore beneficial to keep all dimensions.

Moreover, Table 4 provides the details of the performance according to different test durations on the LRE09 corpus. It can be observed that the hybrid fusion benefits the performance of longer utterance scenarios more than shorter utterance one.

In conclusion, the experiments demonstrate that both feature concatenation and backend-fusion schemes work for either noisy or large-scale dataset. To be more specific, with the proposed hybrid fusion, the system average cost performance C_{avg} decrease from baseline 0.156 to 0.096 (corresponding to a +38.5% gain) for DARPA RATS. On the other hand, for the NIST LRE09 corpus, the performance C_{avg} on the whole test set is decreased from 0.158 to 0.085 (corresponding to a +46.2% gain).

8. DISCUSSION AND CONCLUSION

A multiple feature front-end set combined with various back-end combinations were proposed for a system fusion framework to fully explore robust LID in clean, noisy, and channel mismatch conditions. We considered difficult real-life scenarios for language recognition, where the test utterances were noisy and of varying duration, similar to what has been observed in the challenging DARPA RATS and NIST LRE09 scenarios. To address noise, channel, and duration mismatch, robust front-end processing is an obvious necessity. In this study, we systematically investigated a series of front-ends and back-ends, which demonstrated that by properly fusing various types of acoustic features and back-end classifiers, performance can be improved significantly. In addition, the latest proposed GCDS back-end outperforms a generative Gaussian back-end. To be more specific, for the DAPRA RATS scenario, hybrid fusion benefits average cost function C_{avg} with a relative +38.5% improvement. For the NIST LRE09 relatively clean scenario, the performance of whole utterance achieved a +46.2% relative improvement. These observations offer useful practices for other practitioners in the LID field.

For the next steps, speech enhancement techniques will be investigated. It is noted that speech enhancement (such as non-negative matrix factorization in [39]) can improve the audio quality and therefore should be beneficial for noisy, corrupted data.

9. REFERENCES

[1] N. Dehak, A. McCree, D. Reynolds, F. Richardson, E. Singer, D. Sturim, and P. Torres-Carrasquillo, "MITLL 2011 Language Recognition Evaluation System Description," in *Proc. NIST 2011 Language Recognition Evaluation Workshop*, Atlanta, USA, Dec. 2011.

[2] N. Brummer, S. Cumani, O. Glembek, P. Matejka Karafiat, J. Pesan, O. Plchot, M. Soufifar1, and E. de Villiers, "Brno276 System Description for NIST LRE 2011," in *Proc. NIST 2011 Language Recognition Evaluation Workshop*, Atlanta ,USA, Dec. 2011.

[3] G. Liu, and J. H. L. Hansen, "A systematic strategy for robust automatic dialect identification," in *Proc. EUSIPCO*, Barcelona, Spain, 2011, pp. 2138-2141.

[4] G. Liu, Y. Lei, and J. H. L. Hansen, "Dialect Identification: Impact of Differences between Read versus Spontaneous Speech," in *Proc. EUSIPCO*, Aalborg, Denmark, Aug. 2010, pp. 2003-2006.

[5] G. Liu, S. O. Sadjadi, T. Hasan, J. Suh, C. Zhang, M. Mehrabani, H. Boril, A. Sangwan, and J. H. L. Hansen, "UTD-CRSS systems for NIST language recognition evaluation 2011," in *Proc. NIST 2011 Language Recognition Evaluation Workshop*, Atlanta, USA, Dec. 2011.

[6] W. Shen, et al. "Experiments with lattice-based PPRLM language identification," in *Proc. Odyssey*, Puerto Rico, 2006, pp. 1-6.

[7] Q. Zhang, H. Boril, and J.H.L. Hansen, " Supervector pre-processing for PRSVM-based Chinese and Arabic dialect identification," in *Proc. ICASSP*, Vancouver, Canada, May 2013, pp. 7363-7367.

[8] N. Dehak, P. Dumouchel, and P. Kenny, "Modeling Prosodic Features With Joint Factor Analysis for Speaker Verification," *IEEE Trans. Audio Speech Lang. Process.*, vol.15, no.7, pp. 2095-2103, Sept. 2007.

[9] G. Liu, T. Hasan, H. Boril, and J.H.L. Hansen, "An investigation on back-end for speaker recognition in multi-session enrollment," in *Proc. ICASSP*, Vancouver, Canada, May 2013, pp. 7755-7759.

[10] D. Martinez, L. Burget, L. Ferrer, and N. Scheffer, "iVector-based prosodic system for language identification," in *Proc. ICASSP*, Kyoto,Japan, March 2012, pp. 4861-4864.

[11] Y. Lei, et al., "The CRSS Systems for the 2010 NIST Speaker Recognition Evaluation," NIST 2010 Speaker Recognition Evaluation Workshop, Brno, Czech Republic, 24-25 Jun. 2010.

[12] C. Yu, G. Liu, S. Hahm, and J.H.L. Hansen, "Uncertainty Propagation in Front End Factor Analysis For Noise Robust Speaker Recognition," in Proc. ICASSP, Florence, Italy, May 2014.

[13] J. Suh, et al., "Exploring Hilbert envelope based acoustic features in i-vector speaker verification using HT-PLDA", SRE2011 Workshop, Atlanta, USA.

[14] N. Brummer, et. al., "ABC System description for NIST SRE 2012," in Proc. NIST Speaker Recognition Evaluation, Orlando, FL, USA, Dec. 2012.

[15] L. Ferrer, M. McLaren, N. Scheffer, Y. Lei, M. Graciarena, and V. Mitra, "A Noise-Robust System for NIST 2012 Speaker Recognition Evaluation," in Proc. Interspeech, Lyon, France, Aug. 2013, pp. 1981-1985.

[16] U. Yapanel, and J.H.L Hansen. "A new perceptually motivated MVDR-based acoustic front-end (PMVDR) for robust automatic speech recognition," *Speech Communication*, vol. 50, no. 2, pp.142-152, 2008.

[17] Y. Shao, S Srinivasan, and D. Wang. "Incorporating auditory feature uncertainties in robust speaker identification," in *Proc. ICASSP*, Honolulu, USA, 2007, vol. 4, pp. 277-280.

[18] M. A. Kohler, and M. Kennedy. "Language identification using shifted delta cepstra," in *Proc. 45th Midwest Symposium on Circuits and Systems*, MWSCAS, 2002, vol. 3, pp. 69-72.

[19] X. Zhou, et al. "Linear versus mel frequency cepstral coefficients for speaker recognition," in *IEEE Automatic Speech Recognition and Understanding workshop*, 2011.

[20] T Kinnunen, et al. "Low-Variance Multitaper MFCC Features: A Case Study in Robust Speaker Verification," *IEEE Trans. Audio Speech Lang. Process.*, vol. 20, no.7, pp. 1990-2001, 2012.

[21] C. Kim, and R.M. Stern. "Power-normalized cepstral coefficients (pncc) for robust speech recognition," in *Proc. ICASSP*, Kyoto, Japan, March 2012, pp. 4101-4104.

[22] H. Hermansky, and N. Morgan. "RASTA processing of speech," *IEEE Trans. Audio Speech Lang. Process.*, vol. 2, no. 4, pp. 578-589, 1994

[23] T. Hasan, S. O. Sadjadi, G. Liu, N. Shokouhi, H. Boril, and J. H. L. Hansen, "CRSS systems for 2012 NIST speaker recognition evaluation," in *Proc. ICASSP*, Vancouver, Canada, May 2013, pp. 6783-6787.

[24] T. Rahman, S. Mariooryad, S. Keshavamurthy, G. Liu, J. H. L. Hansen, and C. Busso, "Detecting sleepiness by fusing classifiers trained with novel acoustic features," in *Proc. Interspeech*, Florence, Italy, Aug. 2011, pp. 3285-3288.

[25] G. Liu, Y. Lei, and J. H. L. Hansen, "A novel feature extraction strategy for multi-stream robust emotion identification," in *Proc. Interspeech*, Makuhari, Japan, Sep. 2010, pp. 26-30.

[26] P.A. Torres-Carrasquillo, et al. "The MITLL NIST LRE 2009 language recognition system." in *Proc. ICASSP*, Dallas, USA, March 2010, pp. 4994-4997.

[27] M. Penagarikano, A. Varona, M. Diez, L. Javier Rodriguez-Fuentes, and G. Borde, "Study of different backends in a state-of-the-art language recognition system," in *Proc. Interspeech*, 2012, Portland, USA, pp. 2049-2052.

[28] V. Hautamaki, K. A. Lee, D. v. Leeuwen, R. Saeidi, A. Larcher, T. Kinnunen, T. Hasan, S. O. Sadjadi, G. Liu, H. Boril, J.H.L. Hansen and B. Fauve, "Automatic regularization of cross-entropy cost for speaker recognition fusion," in *Proc. Interspeech*, Lyon, France, Aug. 2013.

[29] R. Saeidi, et al. , "I4U submission to NIST SRE 2012: A large-scale collaborative effort for noise-robust speaker verification," in *Proc. Interspeech*, Lyon, France, Aug. 2013.

[30] N. Brümmer, "Focal multi-class—tools for evaluation, calibration and fusion of, and decision-making with, multi-class statistical pattern recognition scores," [Online]. Available: http://sites.google.com/site/nikobrummer/focalmulticlass, Jun. 2007.

[31] G. Liu, C. Zhang, and J. H.L. Hansen, "A Linguistic Data Acquisition Front-End for Language Recognition Evaluation," in *Proc. Odyssey*, Singapore, Jun. 2012.

[32] L. N. Tan, and A. Alwan, "Multi-band summary correlogram-based pitch detection for noisy speech", *Speech Communication*, vol.55, pp. 841-856, Sep. 2013.

[33] D. Martinez, O. Plchot, L. Burget, O. Glembek, and P. Matejka, "Language recognition in iVectors space," in *Proc. Interspeech*, Florence, Italy, Sept. 2011, pp. 861–864.

[34] N. Dehak, P.A. Torres-Carrasquillo, D. Reynolds, and R. Dehak, "Language recognition via ivectors and dimensionality reduction," in *Proc. Interspeech*, Florence, Italy, Sept. 2011, pp.857–860.

[35] P. Kenny, G. Boulianne, and P. Dumouchel, "Eigenvoice modeling with sparse training data," *IEEE Trans. Speech Audio Process.*, vol. 13, pp. 345–354, May 2005.

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

[37] D. Garcia-Romero and C. Y. Espy-Wilson, "Analysis of I-vector Length Normalization in Speaker Recognition Systems," in Proc. Interspeech, Florence, Italy, 2011, pp. 249-252.

[38] "The 2009 NIST Language Recognition Evaluation Plan (LRE09)", [Online] Available: http://www.itl.nist.gov/iad/mig//tests/lang/2009.

[39] G. Liu, D. Dimitriadis and E. Bocchieri, "Robust speech enhancement techniques for ASR in non-stationary noise and dynamic environments", in Proc. Interspeech, Lyon, France, Aug. 2013, pp. 3017-3021.