Impact of Noise Reduction and Spectrum Estimation on Noise Robust Speaker Identification

Keith W. Godin, Seyed Omid Sadjadi, John H. L. Hansen⋆

Center for Robust Speech Systems (CRSS)
The University of Texas at Dallas, Richardson, TX 75080-3021, U.S.A.
godin@ieee.org, {sadjadi, john.hansen}@utdallas.edu

Abstract

Many spectrum estimation methods and speech enhancement algorithms have previously been evaluated for noise-robust speaker identification (SID). However, these techniques have mostly been evaluated over artificially noised, mismatched training tasks with GMM-UBM speaker models. It is therefore unclear whether performance improvements observed with these methods translate to a broader range of noisy SID tasks. This study compares selected spectrum estimation methods from three classes: cochlear filterbanks, alternative time-domain windowing, and linear prediction-based techniques, as well as a set of frequency-domain noise reduction algorithms, across a suite of 8 evaluation tasks. The evaluation tasks are designed to expand upon the limited tasks addressed in past evaluations by exploring three research questions: performance on real noise versus artificial noise, performance on matched training tasks versus mismatched tasks, and performance when paired with an i-vector backend versus a GMM-UBM backend. We find that noise-robust spectrum estimation methods can improve the performance of SID systems over the range of noise tasks evaluated, including real noisy tasks, matched training tasks, and i-vector backends. However, performance on the typical GMM-UBM mismatched artificially noised case did not predict performance on other tasks. Finally, the matched enrollment case is a significantly different problem than the mismatched enrollment case.

Index Terms: mismatched condition, noise robustness, robust features, speaker identification, speech enhancement

1. Introduction

Robustness to environmental noise is essential to most practical applications of speaker identification systems. A significant body of work exists exploring spectrum estimation and noise reduction algorithms for this purpose, e.g. [1–21] . However, review of this literature does not reveal which methods are most successful or what their characteristics are across different noise types or conditions nor how the algorithms compare to one another. To this end, [20] carried out an extensive comparison of several spectrum estimation methods. They found that selection of the spectrum estimation technique had a significant impact on SID performance in noise, and that the best spectrum estimator depended on noise type and level.

Surveying the literature on spectrum estimation for noise robust SID, we find three general classes of modifications to the baseline FFT-based MFCCs. One class of algorithms is based on linear prediction. [1] compared objective functions for linear prediction coefficient estimation. [12] and [14] discuss weighted linear prediction. [19] discusses regularized variants of linear prediction methods. Finally, [18] investigates frequency domain linear prediction. This class of algorithms was evaluated in [20], where they found that, at all noise levels, the unweighted linear prediction offered performance improvement over baseline FFT-MFCCs, while further improvements using weighted linear prediction methods were possible, depending on noise type and level.

A second class of alternative spectrum estimation methods depend on a different window. [10, 21] discuss multi-taper spectrum estimation, which uses the average of spectrum estimates computed from a suite of windows applied to each frame. [16] demonstrated improvements using the asymmetric G.729 speech coding window. The comparative evaluation of [20] found that alternative windows offered performance improvements over baseline for conditions with SNRs greater than or equal to 10 dB, but were less effective than the linear prediction methods for all but the original clean condition.

Finally, a third class of alternative spectrum estimation methods uses a gammatone filter bank as a cochlear filter simulator. [8] demonstrated features computed from a 128 channel gammatone filter bank, post processed with downsampling, cube-root, and DCT. [13] instead postprocessed the sub-band energies from a 32 channel filter bank by taking the frame-mean of the Hilbert envelope of the filter bank outputs. [9] applied a tensor cross-speaker normalization to the cochlear filter bank outputs. [17] applied an energy-based mask to remove non-speech time-frequency components. Cochlear filter bank methods were not included in [20].

In addition to spectrum estimation methods, speech enhancement has been applied as a noise reduction method for SID. [2, 3, 7, 13, 14] found significant performance improvement using power spectral subtraction (PSS). [5] discussed a Gaussian Mixture Model (GMM) based speech enhancement method. [6] investigated the relative autocorrelation spectrum (RAS), an autocorrelation domain technique. Finally, [13] evaluated four speech enhancement algorithms, showing that the best technique depended on noise type and level. The literature does not include a comparative SID evaluation that includes both speech enhancement methods and spectrum estimation methods.

In addition to the comparison of cochlear-filter class spectrum estimation and noise reduction methods for noise robust SID, three additional questions remain. One is the performance impact of the definition of the noise robustness problem. A SID system depends on three partitions of data: development data to estimate system models, enrollment data to estimate speaker models, and evaluation data on which to perform recognition. Noise robustness tasks are defined in terms of including or not including noise in any combination of these partitions. Among

⋆This project was funded 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.

[1–21], noise is usually included only in the evaluation data. In the following evaluation we will refer to as the 'mismatched' case. It is worth considering two additional noise robustness problems: in one, all three of the data partitions are affected by the same source of additive noise. We will refer to this as the 'matched' case. In the other, general characteristics of the noise affecting the evaluation data are known *a priori*, and are used to generate additional sets of artificially noised development and enrollment data, which are used in addition to the clean source data. This occurred, for example, in the most recent NIST Speaker Recognition Evaluation, SRE-2012. We will refer to this as the 'mixed' case. [20] considers only the mismatched case. The literature does not include a comparative evaluation of noise reduction methods and spectrum estimation methods on matched-case and mixed-case noisy SID, which may prove to be significantly different problems from the mismatched case.

The literature also does not offer evidence on whether the impact of noise reduction and spectrum estimation methods depends on the pattern recognition system used. The GMM-UBM method [22] is most common in the related literature, while the current state-of-the-art technique is the i-vector method. Robustness improvements due to spectrum estimation or noise reduction methods may not carry over to an i-vector system.

Finally, most of the evaluations of [1–21] used artificially noised speech segments. Generating noisy speech by adding recordings of environmental noise to a clean speech recording has the advantage that the SNR can be controlled. It is also less expensive to collect and may be based on readily available clean speech corpora. Unfortunately, naturally noised SID corpora have not been widely available until the recent release of the SRE-2012 evaluation data. Artificially noised segments do not include the Lombard effect inherent to speech production in noise (the effect of this on SID is discussed in [23]), and the noise signals used are statistically homogeneous. Therefore, experimental results from the literature may not carry over to the real-noise case.

It is the purpose of this study to broaden the classes of spectrum estimation techniques under investigation, to include noise reduction methods in a comparison to spectrum estimation techniques, and to significantly increase the scope of the evaluation to answer new research questions regarding the role of noise robustness problem definition, impact on i-vector backends, and effect of real noise as compared to artificial noise. In order to carry out this evaluation, a set of four evaluation tasks is described in the next section.

2. Evaluation Tasks

In order to investigate the three primary research questions, GMM-UBM system and i-vector systems are used with each of the front-end techniques on four SID tasks, resulting in 8 equal error-rate (EER) measurements per technique. These four tasks cover the four combinations of real and artificial noise, crossed with matched and mismatched evaluation conditions. Table 1 shows which evaluation tasks are used to fulfill each evaluation function.

Table 1: Speaker ID Evaluation tasks to address noise type and evaluation regime research questions.

Matched Train	Real/Art.	Corpus
Matd.	Real	NoTel
Matd.	Art.	Art. Noised SRE10
MisMd.	Real	SRE12 Cond. 5
MisMd.	Art.	SRE12 Cond. 4

The Noisy Telephone (NoTel)[1] corpus is used for the matched condition real-noise evaluation. NoTel is a set of 11,027 sessions of telephone speech recorded in various naturally noisy conditions, including roadside and in-vehicle. From this data set, we have developed speaker verification task of 92,000 trials (3,000 target trials), with 300 enrolled speakers. The task developed for this evaluation from the NoTel corpus is a 'matched' task in a broad sense, in that the development, enrollment, and evaluation data are all noisy. However, as the NoTel corpus contains several noise types, speakers may be heard in a different noise type in enrollment versus evaluation.

An artificially noised version of the SRE-2010 [24] male and female telephone-telephone trials is used for the matched condition artificial-noise evaluation. Recordings of HVAC and large crowd noise (LCN) have been added to the development, enrollment, and evaluation data at 6 and 15 dB-P SNR. db-P refers to the psophometric weighting that has been used to calculate the speech and noise powers in the noise adding process, in the same manner as the SRE-2012 artificially noised segments[2]. The noise type and level are selected randomly for each file. The development data has been drawn from SRE 2005, 2006, and 2008, and comprises 9,120 files. As with NoTel, this task is matched in a broad sense, as enrollment and evaluation might occur under different noise type and level regimes.

Finally, Common Conditions 4 and 5 of the SRE-2012 task are used for the mismatched naturally and artificially noised evaluation conditions. Common Condition 5 of SRE-2012 has trials with naturally noised evaluation data, as reported by the speakers under collect, and comprises 73,008 trials, with 1,719 target trials. Common Condition 4 has trials with artificially noised evaluation data and comprises 143,727 trials, with 3,105 target trials. For this task, development data is drawn from SRE 2004, 2005, 2006, 2008, and 2010 and is comprised of 8,401 files.

3. Speaker ID System

For the evaluation, two speaker modeling backends are used: a GMM-UBM architecture [22] and an i-vector architecture [26]. For each task, these two backends share a 512-mixture UBM with diagonal covariance matrices. The GMM-UBM system uses MAP adapted speaker models with a relevance factor of 4. The i-vector system uses a total variability (TV) matrix estimated over 5 EM iterations. 400-dimensional i-vectors are reduced to 250 dimensions with LDA. A Gaussian PLDA model of 250 columns in the Eigenvoice matrix is used for scoring . All systems are gender independent. This is required in the case of NoTel, which does not include gender labels, and has been carried through to the other systems for consistency and to reduce the computational burden.

The baseline frontend uses the unsupervised Combo-SAD [27], no pre-emphasis, 17 dimensional MFCCs computed from a 17-filter filter bank over 300-3400 Hz. The baseline and all other frontends are post-processed with cepstral mean subtraction over a 3-second sliding window, and appended with delta and acceleration coefficients. The baseline system EERs are shown in Table 2. Increasing the UBM size to 1024 or 2048 mixtures may reduce the baseline EER by one or two percentage points, but has been reduced to 512 mixtures to reduce the computational burden of the evaluation. For the same reason, the system development data has been reduced by half of the amount otherwise available

[1] The Noisy Telephone corpus is provided by AFRL and is not publicly available.

[2] A MATLAB script for filtering and noise adding based on the psophometric weighting is available from [25].

Table 2: Baseline system performance

Corpus	Backend	Noise	Enroll	EER (%)
NoTel	i-vector	Real	Mtch.	4.967
SRE10-N	i-vector	Art.	Mtch.	7.602
SRE12-4	i-vector	Real	MisMtch.	10.413
SRE12-5	i-vector	Art.	MisMtch.	12.786
NoTel	GMM-UBM	Real	Mtch.	11.250
SRE10-N	GMM-UBM	Art.	Mtch.	33.192
SRE12-4	GMM-UBM	Real	MisMtch.	14.078
SRE12-5	GMM-UBM	Art.	MisMtch.	20.160

given the source corpora.

4. Methods Under Evaluation

4.1. Linear Prediction Spectrum Estimation

Linear prediction (LP) based features were among the earliest acoustic features used for SID [28]. The cepstral recursion was the original cepstral linear prediction feature [29]. More recently, the linear prediction power spectrum estimate has been substituted for the FFT spectrum estimate in MFCCs, yielding LP-MFCCs [1, 4, 12, 14, 18, 19]. Also, noise robust features have been based on the Minimum Variance Distortionless Response (MVDR) spectrum estimate, which is based on the linear prediction spectrum estimate [30]. A perceptually warped version (PMVDR) has shown improvement on SID tasks [20, 31, 32].

Improvements to LP-MFCCs may result from weighting, stabilization, and regularization. [12, 14] evaluated a suite of weighted and stabilized LP estimators, finding that the methods generally improved over LP-MFCCs and FFT-MFCCs, and that stabilized weighted linear prediction offered the most consistent improvements in EER. [19] evaluated regularized LP estimators, finding that regularization improved each variant of weighted LP estimator. Finally, frequency domain linear prediction (FDLP) has also been shown to improve the performance of SID systems in noise [18]. Based on these results, we have chosen 20th-order LP-MCCs, 20th-order regularized LP-MFCCs (RLP), FDLP, and PMVDR as representative of this class. The RLP code is contained in [19], except modified to scale the spectrum estimate by the RMS of the prediction residual. We have used $\lambda = 10^{-7}$. The FDLP implementation is from [33], modified to remove cepstral liftering and gain normalization, use 10 s segments, and 30th order LP models for each subband, to match the parameter settings described in [18].

4.2. Cochlear Filter Bank Spectrum Estimation

A class of features applied to SID noise robustness relies on a bank of gammatone filters to provide the spectrum estimate. These features often incorporate design choices motivated by models of the human auditory system. The features in this class differ in the way that the spectrum estimate is processed into feature frames. For this evaluation, Mean Hilbert Envelope Coefficients (MHECs) have been selected as representative of the class. The MHEC implementation used here is described in [34].

4.3. Alternative Window Spectrum Estimation

Alternative windowing schemes may also improve the noise robustness of SID systems. [16] investigated two asymmetric windows that improved performance versus the baseline Hamming window. The present evaluation includes the G.729 speech coder window evaluated in [16].

Another method to reduce the spectrum estimate variance is to average the estimate over multiple windows. [21, 35] discuss these windows, while [10, 21] demonstrated their effectiveness for SID noise robustness. Based on the results of that study, the present evaluation includes the Sine Weighted Cepstrum Estimator (SWCE) set of tapers. The implementation is available from [36].

4.4. Noise Reduction

We have selected three speech enhancement methods as representative of those that have been applied to SID: power spectral subtraction (PSS) [37], Wiener filtering (WF) [38], and log-MMSE enhancement [39]. The *a priori* SNR used for our PSS implementation is defined as:

$$\xi_{ML} = \max\left(\frac{\Gamma_{xx}}{\Gamma_{nn}} - 1, \xi_{ML}^{min}\right), \tag{1}$$

where Γ_{xx} is the observation power spectrum and Γ_{nn} is the estimated noise power spectrum. Without this flooring, PSS significantly degrades SID performance. We have used $\xi_{ML}^{min} = 0.1$, i.e. maximum 10 dB attenuation. Both components of the decision-directed (DD) *a priori* SNR [38] used for the Wiener filter and log-MMSE have similarly been floored. The gain functions used for PSS and Wiener filter are:

$$G_{PSS} = \sqrt{\frac{\xi_{ML}}{\gamma}} \tag{2}$$

$$G_{WF} = \frac{\xi_{DD}}{\xi_{DD} + 1} \tag{3}$$

where γ is the *a posteriori* SNR:

$$\gamma = \frac{\Gamma_{xx}}{\Gamma_{nn}} \tag{4}$$

The log-MMSE gain function is as given in [39]. All three enhancement implementations rely on the minimum statistics noise estimator [40] as implemented in [41].

5. Evaluation Results

5.1. Raw Performance Impact

In this study, relative improvement or degradation in excess of 5% is treated as significant. Table 3 shows the relative improvement (or degradation) versus baseline for each method under each of the 8 evaluation regimes. Relative improvement greater than 5% is marked in green, with relative degradation greater than 5% in red. PSS, log-MMSE, MHEC, SWCE, and PMVDR resulted in significant relative improvements averaged across all tasks. Wiener filter and RLP resulted in significant degradation, averaged across all tasks. The G.729 window, LP-MFCCs, and FDLP did not result in significant shifts when averaged across all tasks.

5.2. Influence of Evaluation Factors

Table 3 shows that there are significant differences in results across tasks for each method. Three evaluation factors influence the results. All of the methods improved more or degraded less on real noise than on artificial noise (Table 4). However, matched data reduced or eliminated the gains seen in mismatched tasks for the enhancement methods, as well as for MHEC, PMVDR, and FDLP spectrum estimation methods (Table 5). Negligible effects of matched data were observed for the G.729 window and LP-MFCCs. Finally, MHECs, SWCE, and PMVDR performed significantly better when paired with an i-vector backend (Table 6), while the Wiener filter performed significantly worse with an i-vector backend.

Table 3: Relative % improvement (degradation) of methods to MFCC baseline. Columns are test conditions, in System-Noise-Enrollment format. Iv are i-vector systems, GU are GMM-UBM systems. R for real noise, Ar for artificial noise, M for matched training condition, MM for mismatched training condition. Green numbers denote improvement in excess of 5% relative. Red numbers denote degradation in excess of 5% relative.

Method	Iv-R-M	Iv-Ar-M	Iv-R-MM	Iv-Ar-MM	GU-R-M	GU-Ar-M	GU-R-MM	GU-Ar-MM	Mean
PSS	(4.02)	(7.75)	+25.1	+9.58	(1.62)	0.00	+24.0	+3.87	+6.15
Wiener Filt.	(18.8)	(83.9)	+20.7	(80.5)	+0.155	+3.92	+29.0	+0.208	(16.2)
Log-MMSE	(8E-3)	+0.484	+22.3	+10.6	+0.839	+1.74	+19.8	+20.6	+9.55
MHEC	+14.8	(5.91)	+40.5	+24.4	+19.2	(3.4)	+6.61	+11.1	+13.4
G.729 Win.	+4.04	(1.33)	+2.80	+0.508	+10.8	(2.55)	(0.824)	(0.283)	+1.65
SWCE Win.	+14.8	+12.7	+8.94	(1.94)	+11.5	(1.99)	+0.419	+3.06	+5.93
PMVDR	+20.1	+3.39	+43.0	+27.5	+26.5	(9.33)	(4.13)	(3.84)	+12.9
RLP	(30.2)	(142)	(146)	(156)	(29.2)	(14.0)	(56.2)	(43.9)	77.2
LP-MFCC	+2.69	(7.75)	+6.15	+4.03	+12.4	(4.25)	(13.0)	+1.28	+0.185
FDLP	(2.00)	(5.61)	+21.3	+11.6	+6.73	+1.75	+8.27	10.0	+4.00

Table 4: Average relative improvement (degradation) across real and artificially noised evaluation regimes. Positive numbers in the 'Shift' column suggest greater relative improvement on real data than on artificially noised data.

Method	Artificial	Real	Shift
PSS	+1.43	+10.9	+9.45
Wiener Filt.	(40.1)	+7.77	+47.8
Log-MMSE	+8.35	+10.8	+2.40
MHEC	+6.54	+20.3	+13.7
G.729 Win.	(0.914)	+4.21	+5.12
SWCE Win.	+2.95	+8.91	+5.95
PMVDR	+4.42	+21.4	+17.0
RLP	(89.0)	(65.4)	+23.6
LP-MFCC	(1.68)	+2.05	+3.72
FDLP	(0.574)	+8.57	+9.14

Table 5: Average relative improvement (degradation) across matched and mismatched evaluation regimes. Positive numbers in the 'Shift' column suggest greater relative improvement on matched evaluation tasks than on mismatched evaluation tasks.

Method	Mismatched	Matched	Shift
PSS	+15.6	(3.35)	(19.0)
Wiener Filt.	(7.65)	(24.7)	(17.0)
Log-MMSE	+18.3	+0.764	(17.6)
MHEC	+20.6	+6.17	(14.5)
G.729 Win.	+0.549	+2.74	+2.20
SWCE Win.	+3.25	+9.24	+6.62
PMVDR	+15.6	+10.2	(5.45)
RLP	(101)	(53.7)	+46.9
LP-MFCC	(0.400)	+0.771	+1.17
FDLP	+7.78	+0.219	(7.56)

Table 6: Average relative improvement (degradation) across GMM-UBM and i-vector evaluation regimes. Positive numbers in the 'Shift' column suggest greater relative improvement with an i-vector backend than with a GMM-UBM backend.

Method	GMM-UBM	i-vect.	Shift
PSS	+6.56	+5.74	(0.816)
Wiener Filt.	+8.33	(40.64)	(49.0)
Log-MMSE	+10.8	+8.35	(2.41)
MHEC	+8.37	+18.4	+10.1
G.729 Win.	+1.79	+1.50	(0.288)
SWCE Win.	+3.25	+8.61	+5.37
PMVDR	+2.31	+23.5	+21.2
RLP	(35.8)	(119)	(82.7)
LP-MFCC	(0.908)	+1.28	+2.19
FDLP	+1.68	+6.32	+4.64

6. Conclusions

In comparing real noised tasks to artificially noised tasks, relative performance improvements were generally greater for real noised tasks. Given the evaluation tasks used here, these results suggest that artificially noised SID tasks do not predict performance gains or losses on real-noised tasks due to spectrum estimation and noise reduction techniques. Artificially noised corpora have been important to research efforts because they can be controlled for noise level and are simple to create. However, they are only useful to the extent to which they predict performance on real-noise tasks, and given the results here, further evaluation should investigate under which conditions artifically noised tasks may or may not predict performance on real-noised tasks.

For some methods, approximately the same impact on performance was observed for GMM-UBM backends as for i-vector backends, while for others, significantly more improvement was observed when paired with an i-vector backend. These results suggest that moving to an i-vector backend does not necessarily erase performance gains observed in the GMM-UBM case, and that performance gains or losses observed with a GMM-UBM backend do not predict gains or losses observed with an i-vector backend. GMM-UBM techniques are important because they are simpler and more accessible, and offer a link to almost two decades of SID research. However, i-vector techniques are the state-of-the-art in SID. Research in noise robust SID should not ignore i-vectors and other advanced backends.

The methods evaluated here generally offered significantly less improvement, or significantly more degradation, on matched enrollment tasks than on mismatched tasks. The exception is the alternative window methods: The SWCE method did not offer significant performance improvement on the mismatched tasks, but did offer significant improvement on the matched tasks. We conclude that matched tasks are a significantly different problem in SID than mismatched tasks, and that only a limited set of extant noise robust frontend techniques may offer performance gains on matched tasks. Research efforts in noise robust SID have not yet addressed the matched enrollment case.

Finally, [16, 18] are examples of recent research efforts that have treated FFT-MFCCs as the baseline technique. For mismatched tasks, PSS or log-MMSE should be considered part of a noise robust SID baseline. Further, fair comparison demands that multi-taper and cochlear spectrum estimation, as well as MVDR, be considered in addition to FFT-MFCCs in any evaluation of noise robust SID frontends.

7. References

[1] R. P. Ramachandran, M. S. Zilovic, and R. J. Mammone, "A comparative study of robust linear predictive analysis methods with applications to speaker identification," *IEEE Trans. Speech Audio Process.*, vol. 3, pp. 117–125, 1995.

[2] J. Ortega-Garcia and J. Gonzalez-Rodriguez, "Overview of speech enhancement techniques for automatic speaker recognition," in *Proc. ICSLP*, 1996, pp. 929–932.

[3] A. Drygajlo and M. El-Maliki, "Speaker verification in noisy environments with combined spectral subtraction and missing feature theory," in *Proc. IEEE ICASSP*, 1998, pp. 121–124.

[4] M. S. Zilovic, R. P. Ramachandran, and R. J. Mammone, "Speaker identification based on the use of robust cepstral features obtained from pole-zero transfer functions," *IEEE Trans. Speech Audio Process.*, vol. 6, pp. 260–267, 1998.

[5] T. Ganchev, I. Potamitis, N. Fakotakis, and G. Kokkinakis, "Text-independent speaker verification for real fast-varying noisy environments," *Int. J. Speech Technol.*, vol. 7, pp. 281–292, 2004.

[6] K.-H. Yuo, T.-H. Hwang, and H.-C. Wang, "Combination of autocorrelation-based features and projection measure technique for speaker identification," *IEEE Trans. Speech Audio Process.*, vol. 13, pp. 565–574, 2005.

[7] A. Moreno-Daniel, J. Nolazco-Flores, T. Wada, and B.-H. Juang, "Acoustic model enhancement: An adaptation technique for speaker verification under noisy environments," in *Proc. IEEE ICASSP*, 2007, pp. 289–292.

[8] Y. Shao, S. Srinivasan, and D. Wang, "Incorporating auditory feature uncertainties in robust speaker identification," in *Proc. IEEE ICASSP*, 2007, pp. 277–280.

[9] Q. Wu and L. Zhang, "Auditory sparse representation for robust speaker recognition based on tensor structure," *EURASIP J. Audio Speech and Music Process.*, vol. 2008, pp. 1–9, 2008.

[10] T. Kinnunen, R. Saeidi, J. Sandberg, and M. Hansson-Sandsten, "What else is new than the hamming window? robust mfccs for speaker recognition via multitapering," in *Proc. INTERSPEECH*, 2010, pp. 2734–2737.

[11] Q. Li and Y. Huang, "Robust speaker identification using an auditory-based feature," in *Proc. IEEE ICASSP*, 2010, pp. 4514–4517.

[12] J. Pohjalainen, R. Saeidi, T. Kinnunen, and P. Alku, "Extended weighted linear prediction (XLP) analysis of speech and its application to speaker verification in adverse conditions," in *Proc. INTERSPEECH*, 2010, pp. 1477–1480.

[13] S. O. Sadjadi and J. H. L. Hansen, "Assessment of single-channel speech enhancement techniques for speaker identification under mismatched conditions," in *Proc. INTERSPEECH*, 2010, pp. 2138–2141.

[14] R. Saeidi, J. Pohjalainen, T. Kinnunen, and P. Alku, "Temporally weighted linear prediction features for tackling additive noise in speaker verification," *IEEE Signal Process. Lett.*, vol. 17, pp. 599–602, 2010.

[15] C. wa Maina, "Approximate Bayesian inference for robust speech processing," Ph.D. dissertation, Drexel Univ., June 2011.

[16] M. J. Alam, P. Kenny, and D. O'Shaughnessy, "On the use of asymmetric-shaped tapers for speaker verification using i-vectors," in *Proc. Odyssey 2012*, 2012.

[17] T.-S. Chi, T.-H. Lin, and C.-C. Hsu, "Spectro-temporal modulation energy based mask for robust speaker identification," *J. Acoust. Soc. Am.*, vol. 131, pp. EL368–EL374, 2012.

[18] S. Ganapathy, S. Thomas, and H. Hermansky, "Feature extraction using 2-D autoregressive models for speaker recognition," in *Proc. Odyssey 2012*, 2012.

[19] C. Hanilci, T. Kinnunen, R. Saeidi, J. Pohjalainen, P. Alku, and F. Ertas, "Regularization of all-pole models for speaker verification under additive noise," in *Proc. Odyssey 2012*, 2012.

[20] C. Hanilci, T. Kinnunen, R. Saeidi, J. Pohjalainen, P. Alku, F. Ertas, J. Sandberg, and M. Hansson-Sandsten, "Comparing spectrum estimators in speaker verification under additive noise degradation," in *Proc. IEEE ICASSP*, 2012, pp. 4769–4772.

[21] T. Kinnunen, R. Saeidi, F. Sedlak, K. A. Lee, J. Sandberg, M. Hansson-Sandsten, and H. Li, "Low-variance multitaper MFCC features: A case study in robust speaker verification," *IEEE Trans. Audio Speech Lang. Process.*, vol. 20, pp. 1990–2001, 2012.

[22] D. A. Reynolds and R. C. Rose, "Robust text-independent speaker identification using Gaussian mixture speaker models," *IEEE Trans. Speech Audio Process.*, vol. 3, pp. 72–83, Jan. 1995.

[23] J. H. L. Hansen and V. Varadarajan, "Analysis and compensation of Lombard speech across noise type and levels with application to in-set/out-of-set speaker recognition," *IEEE Trans. Audio Speech Lang. Process.*, vol. 17, pp. 366–378, 2009.

[24] NIST, "The NIST year 2010 speaker recognition evaluation plan," Natl. Inst. of Standards and Tech. (NIST), Natl. Inst. of Standards and Tech. (NIST), Tech. Rep., 2010.

[25] MATLAB code for Filtering and Noise Adding. [Online]. Available: http://www.utdallas.edu/~sadjadi/AddNoisePSO.m

[26] N. Dehak, Z. N. Karam, D. A. Reynolds, R. Dehak, W. M. Campbell, and J. R. Glass, "A channel-blind system for speaker verification," in *Proc. IEEE ICASSP*, 2011, pp. 4536–4539.

[27] S. O. Sadjadi and J. H. L. Hansen, "Unsupervised speech activity detection using voicing measures and perceptual spectral flux," *IEEE Signal Process. Lett.*, vol. 20, pp. 197–200, 2013.

[28] B. S. Atal, "Automatic recognition of speakers from their voices," *Proc. of the IEEE*, vol. 64, pp. 460–475, Apr. 1976.

[29] J. R. Deller, J. H. L. Hansen, and J. G. Proakis, *Discrete-Time Processing of Speech Signals*. IEEE Press, Piscataway, NJ, 2000, p. pg. 442.

[30] M. N. Murthi and B. D. Rao, "Minimum variance distortionless response (MVDR) modeling of voiced speech," in *Proc. IEEE ICASSP*, 1997, pp. 1687–1690.

[31] U. H. Yapanel and J. H. L. Hansen, "A new perceptually motivated MVDR-based acoustic front-end (PMVDR) for robust automatic speech recognition," *Speech Commun.*, vol. 50, pp. 142–152, 2008.

[32] A. Lawson, P. Vabishchevich, M. Huggins, P. Ardis, B. Battles, and A. Stauffer, "Survey and evaluation of acoustic features for speaker recognition," in *Proc. IEEE ICASSP*, 2011, pp. 5444–5447.

[33] S. Ganapathy. [Online]. Available: http://old-site.clsp.jhu.edu/~sriram/research/fdlp/feat_extract.tar.gz

[34] S. O. Sadjadi and J. H. L. Hansen, "Hilbert envelope based features for robust speaker identification under reverberant mismatched conditions," in *Proc. IEEE ICASSP*, 2011, pp. 5448–5451.

[35] J. Sandberg, M. Hansson-Sandsten, T. Kinnunen, R. Saeidi, P. Flandrin, and P. Borgnat, "Multitaper estimation of frequency-warped cepstra with application to speaker verification," *IEEE Signal Process. Lett.*, vol. 17, pp. 343–346, 2010.

[36] MATLAB code for Multi-taper Spectrum Estimation. [Online]. Available: http://cs.joensuu.fi/pages/tkinnu/multitaper/multitaperspectrum_functions.zip

[37] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Acoust. Speech Signal Process.*, vol. 27, pp. 113–120, 1979.

[38] P. Scalart and J. V. Filho, "Speech enhancement based on a priori signal to noise estimation," in *Proc. IEEE ICASSP*, 1996, pp. 629–632.

[39] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE Trans. Acoust. Speech Signal Process.*, vol. 33, pp. 443–445, 1985.

[40] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *IEEE Trans. Speech Audio Process.*, vol. 9, pp. 504–512, 2001.

[41] M. Brookes *et al.* VOICEBOX: Speech procesing toolbox for MATLAB. [Online]. Available: http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html