Assessment of Single-Channel Speech Enhancement Techniques for Speaker Identification under Mismatched Conditions

Seyed Omid Sadjadi and John H.L. Hansen[⋆]

Center for Robust Speech Systems (CRSS), The University of Texas at Dallas, USA
{sadjadi, john.hansen}@utdallas.edu

Abstract

It is well known that MFCC based speaker identification (SID) systems easily break down under mismatched train and test conditions. In this study, we report on evaluation of four different single-channel speech enhancement front-ends for robust SID under such conditions. Speech files from the YOHO database are corrupted with four types of noise including babble, car, factory, and white at five SNR levels (0–20 dB), and processed using four speech enhancement techniques representing distinct classes of algorithms: spectral subtraction, statistical model-based, subspace, and Wiener filtering. Both processed and unprocessed files are submitted to a SID system trained on clean data. In addition, a new set of acoustic feature parameters based on Hilbert envelope of gammatone filterbank outputs are proposed and evaluated for SID task. Experimental results indicate that: (i) depending on the noise type and SNR level, the enhancement front-ends may help or hurt SID performance, (ii) the proposed feature significantly achieves higher SID accuracy compared to MFCCs under mismatched conditions.
Index Terms: feature extraction, gammatone filterbank, Hilbert envelope, speaker identification, speech enhancement

1. Introduction

Performance of automatic speaker identification (SID) systems is severely degraded when an acoustic mismatch happens between train and test stages. One such mismatch could occur when the system is trained on data collected under laboratory conditions while test data are acquired in real environments where different noise sources are active (e.g., in a car). These noise sources can mask/obscure useful acoustic cues by which SID systems are learned to identify speakers.

Several compensation techniques have been proposed in the literature to alleviating the adverse effect of environmental noise on performance of SID engines, most of which were first developed for noise-robust automatic speech recognition (ASR). Despite its simplicity, spectral subtraction has been shown to be effective as a pre-processing stage in mitigating the impact of mismatch between training and test due to additive noise [1], [2], although, it was assumed that the noise is stationary or of slowly varying nature. Multichannel (e.g., microphone arrays) speech processing techniques have also been employed to provide robustness for SID systems in the presence of ambient noise [3], [4]. However, this imposes additional hardware requirements and more complexity on SID systems, and is not applicable to cases where only a single-channel signal (e.g., telephone) or prerecorded mono speech data is available. Assuming *a priori* knowledge of statistical model of noise, parallel model combination (PMC) has been successfully applied for noise compensation [5], [6]. Missing feature theory has opened new avenues for noise-robust speech systems including SID, by discarding the unreliable portion of data which is severely corrupted by noise, and taking into account only the reliable data (assuming partial-band noise corruption) for likelihood calculations [2], [7], and [8]. To generalize this for unknown full-band noise corruption, in [9] a combination of multi-condition training and the missing feature theory has been adopted and shown to be superior to the baseline system trained on clean data, albeit at the expense of introducing a more complex system.

Another way of dealing with the mismatched conditions in SID is to design acoustic features that are less affected by background noise. Although originally designed to represent acoustic spaces of different phonemes for ASR, the MFCCs have been the most widely used features for SID tasks, probably because they provide acceptable identification accuracy under matched conditions. Also, this makes it possible to easily integrate SID and ASR systems. Nevertheless, it is well-known that MFCC based systems are susceptible to training and test mismatch, and this has motivated extensive research effort to find more robust acoustic features capable of capturing speaker identity conveyed in the speech signal. In particular, feature parameters obtained from the temporal envelope of speech analyzed using a gammatone filterbank have shown promise for SID tasks under mismatched conditions [8], [10], and [11].

In this study, we consider four different single-channel speech enhancement front-ends for noise compensation in a GMM based SID system [12], under additive noise mismatched conditions. These front-ends represent distinct classes of enhancement algorithms including spectral subtraction, statistical model-based, subspace, and Wiener filtering. Speech and noise materials are obtained from the YOHO [13] and NOISEX-92 [14] databases, respectively. Four noisy test conditions at five different SNR levels (0–20 dB) are employed to carry out experiments. In addition, a new set of acoustic feature parameters based on the Hilbert envelope of gammatone filterbank outputs are proposed and benchmarked against the MFCCs for the SID task under mismatched conditions.

2. Speech enhancement front-ends

Table 1 lists four speech enhancement algorithms employed in this study as pre-processing stage for SID to suppress background noise. These methods represent distinct classes of enhancement algorithms including spectral subtraction (SS) [15], statistical model-based (MMSE) [16], subspace (pKLT) [18], and Wiener filtering [17]. The first three algorithms provide an estimate of clean speech magnitude or power spectrum in the short-time Fourier transform (STFT) domain from the available noisy speech spectrum, given that an estimate of the noise spectrum is available. On the other hand, as a perceptually motivated subspace method, pKLT uses the well-known KL transform in conjunction with an auditory model to decompose the noisy speech vector into a signal-plus-noise, as well as a noise subspace, and the clean speech vector is estimated after

[⋆]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. Hansen. Approved for public release; distribution unlimited.

Table 1: *Single-channel speech enhancement front-ends evaluated for SID under additive noise mismatched conditions.*

Algorithm	Equations/Parameters	Ref
SS	(3, 4) / $\beta = 0.05$	[15]
MMSE	(7, 30, 51) / $q = 0.2$, $\alpha = 0.98$	[16]
Wiener	(4-6, 8) / $\beta = \lambda = 0.98$	[17]
pKLT	(34) / $\nu = 0.08$	[18]

removing the noise subspace. Although all these methods are capable of suppressing background noise, their performance is highly dependent on the accuracy of the noise spectrum estimate. This causes the performance to be a trade-off between noise and speech distortion, and we are interested in investigating the impact of such a trade-off on SID performance under mismatched additive noise conditions.

Also given in the table are equations and their respective parameters from the references used for implementing the enhancement algorithms. Frames of 32 ms duration with 50% overlap are used for speech analysis. An initial estimate of noise statistics are obtained from the first 100 ms of each speech signal, and then updated during the enhancement process from speech-absent regions found by voice activity detection (VAD). For the first three methods, a statistical model-based VAD [19] with a threshold parameter $\eta = 1.2$ is used, while for the pKLT a different VAD approach is taken (for more details see [20]). Other implementation details for all the algorithms are well documented in the corresponding references and literature and used *as is* in this study.

3. Mean Hilbert envelope coefficients

In this section, the procedure for extracting a new set of acoustic feature parameters for robust SID under mismatched conditions is described.

The block diagram of the proposed feature extraction scheme is depicted in Fig. 1. First, preemphasized speech signal $s(n)$ is filtered using a 32-channel gammatone filterbank to simulate the effect of auditory filtering which takes place in the cochlea [21]. The filterbank center frequencies are uniformly spaced on equivalent rectangular bandwidth (ERB) scale between 50 and 4000 Hz (assuming a sampling rate of $F_s = 8$ kHz). Next, the temporal envelope of j^{th} channel output $s(n, j)$ is computed as the magnitude of analytical signal obtained using the Hilbert transform. More specifically, let

$$s_a(n, j) = s(n, j) + i\hat{s}(n, j), \tag{1}$$

denote the analytical signal, where $\hat{s}(n, j)$ is the Hilbert transform of $s(n, j)$, and i is the imaginary unit. The temporal envelope $e(n, j)$ is thus calculated as,

$$e(n, j) = \sqrt{s^2(n, j) + \hat{s}^2(n, j)}\,. \tag{2}$$

$e(n, j)$ is also called the Hilbert envelope of the signal $s(n, j)$. In the next stage, the Hilbert envelope $e(n, j)$ is blocked into non-overlapping frames of 10 ms duration. A Hamming window is applied to each frame to minimize the discontinuities at the edges. To estimate the temporal envelope amplitude in frame t, the sample means are computed as,

$$E(t, j) = \frac{1}{N} \sum_{n=0}^{N-1} w(n) e(n, j)\,, \tag{3}$$

where $w(n)$ denotes the Hamming window and N is the frame size in samples. To compress the dynamic range of the envelope amplitude $E(t, j)$, the natural logarithm is applied. Note that $E_{log}(t, j)$ is a measure of energy in j^{th} channel at time frame t. Furthermore, by the duality between time and frequency, it is also a measure of spectral energy in the center frequency of j^{th} channel. These energy features are normalized in each channel by the mean obtained from a long-term average of the spectral energies in that channel as,

$$E_n(t, j) = \frac{E_{log}(t, j)}{\frac{1}{T}\sum_{t=1}^{T} E_{log}(t, j)}\,, \tag{4}$$

with T being the total number of frames. Finally, the discrete cosine transform (DCT) is applied to decorrelate different feature dimensions. This is important because GMMs with *diagonal* covariance matrices can then be used to model the acoustic space of each speaker (as opposed to *full* covariance matrices). The output is a matrix of 32-dimensional spectral features, entitled mean Hilbert envelope coefficients (MHEC).

Since $E_n(t, j)$'s represent spectral energies in different frequency bands, spectrogram-like representations can be made. Fig. 2 demonstrates sample speech spectra across time frames for a clean signal (left), and the same signal corrupted with car noise at 0 dB (right). Top panels have been obtained from STFT analysis in the linear frequency domain, middle panels from $E_n(t, j)$'s, and bottom panels from STFT analysis and mel-band integration in 27 frequency bins. As seen in the figure, the most striking feature of the new signal representation compared to the representation obtained from mel-band integration is that it is less susceptible to the additive noise, and there is a smaller mismatch between the clean and noisy spectra. Therefore, it is expected that higher SID rates can be achieved using the MHECs under mismatched conditions.

4. Experiments

To simulate different noisy conditions, noise samples obtained from the NOISEX-92 database are artificially added to speech signals. First, the active speech level of clean signals are determined using method B described in [22] (no IRS filtering applied). Next, a random segment of the same length as the speech signal is extracted from the noise recordings, appropriately scaled to reach the desired SNR level, and finally added to the clean signal. This is realized using the Filtering and Noise Adding Tool (FaNT) [23]. Four different noise samples including babble, car, factory, and white Gaussian are considered and added to clean signals at five SNR levels in the range 0–20 dB.

The MFCCs are extracted from frames of 25 ms duration at a frame rate of 100 Hz . Out of 27 filterbank log-energies, the first 12 coefficients are retained after applying the DCT (excluding c_0), and delta features are appended to form a 24-dimensional feature vector for each frame. Cepstral mean normalization (CMN) is applied to provide robustness against possible session variability in speakers. The MHECs are obtained using the procedure described in Section 3. A 62-dimensional feature vector is formed by appending the delta features and excluding the energy term.

Two 64-mixture GMM based SID systems (one per feature type) are trained on all 96 clean enrollment utterances of the first 69 speakers (including 12 female and 57 male speakers) in the YOHO database. An energy-based thresholding algorithm is adopted for silence frame removal. All 40 verification utterances are used for subsequent evaluations.

5. Results

Fig. 3 represents identification rates obtained by the two GMM based SID systems under clean and four different noisy test conditions at 5 SNRs covering 0–20 dB range, and for unprocessed and processed test materials. Identification accuracies for the MFCCs and MHECs under clean matched condition are 99.82%

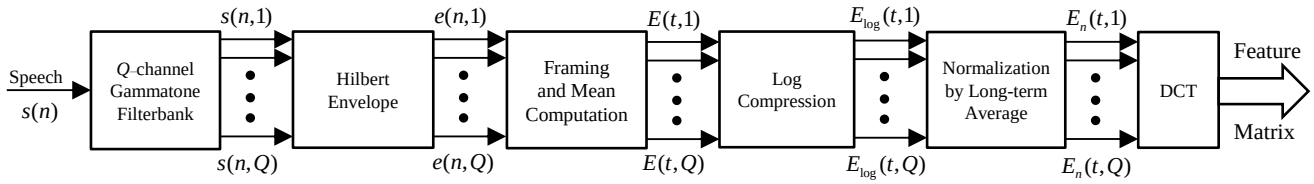

Figure 1: *Block diagram of the proposed feature extraction scheme*

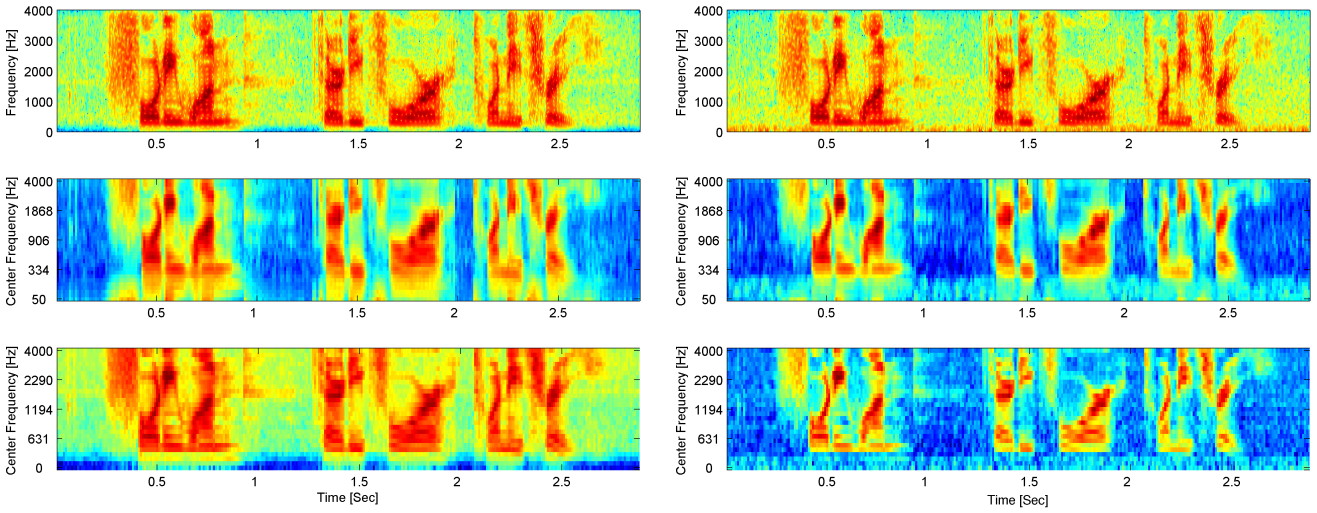

Figure 2: *Sample speech spectrum for: clean signal (the phrase "41-34-23") (left), and signal degraded with car noise at* 0 *dB (right), obtained from: STFT in linear frequency domain (top), mean of the Hilbert envelope of gammatone filterbank outputs (middle), STFT and mel-band integration (bottom).*

and 99.28%, respectively. It is observed that except for the white noise case, the enhancement front-ends provide no improvement in SID accuracy when the SNR is relatively high (15–20 dB). Also, except for the babble noise case, both the SS and MMSE algorithms boost the performance when the SNR is relatively low (10 dB and below). In general, the Wiener and pKLT algorithms performances are the worst. Clearly, the SID system trained on MHECs performs consistently the best across all the mismatched conditions. In particular, while the SID accuracy of the system trained on MFCCs drops dramatically as the SNR level decreases for the test condition under car noise, the system trained on MHECs exhibits almost no decline in identification rate. As discussed in Section 3, this is due to the fact that the MFCCs are easily affected by the additive noise while the MHECs are more robust to a change in the SNR level of the input signal.

Fig. 4 shows the performances of the two systems in clean and four noisy conditions averaged across SNR range 5–20 dB, for the un-processed (UN) as well as the processed test materials using the four enhancement front-ends. It is seen that on average the MHECs yield higher SID accuracy over all the test conditions, and the SS algorithm performs the best in suppressing background noise thus compensating the mismatch between the train and test conditions. Comparing the performance evaluation of different enhancement front-ends in terms of SID accuracy with the evaluation results reported in [20] in terms of intelligibility, reveals that enhancement algorithms that introduce lesser speech distortion provide more improvement in accuracy, (i.e., the SS and MMSE). For instance, in [20] it was shown that the pKLT is very successful in suppressing background noise, however, at the expense of introducing large amounts of signal distortion. Therefore, it can be concluded that the SID systems pay more attention to the signal distortion rather than the noise distortion.

6. Conclusions

The present study considered four distinct speech enhancement front-ends (SS, MMSE, WIN, pKLT) for noise suppression in SID systems under mismatched conditions due to additive noise. It was shown that the spectral subtraction is the most successful method in mitigating the effect of mismatch on the performance of SID systems, especially at relatively low SNR levels (10 dB and below). It was also observed that the SID systems pay more attention to the signal distortion rather than the noise distortion introduced by the front-ends. In addition, a new set of acoustic feature parameters based on the Hilbert envelope of gammatone filterbank outputs was proposed and shown to be superior to the MFCCs in performance under mismatched training and test conditions, while providing almost the same SID accuracy under clean matched conditions.

7. References

[1] J. Ortega-Garca and J. Gonzlez-Rodrguez, "Overview of speech enhancement techniques for automatic speaker recognition," in *Proc. ICSLP'96*, Philadelphia, PA, Oct. 1996, pp. 929–932.

[2] A. Drygajlo and M. El-Maliki,"Speaker verification in noisy environments with combined spectral subtraction and missing feature theory," in *Proc. IEEE ICASSP'98*, Seattle, WA, May 1998, vol. 1, pp. 121–124.

[3] Q. Lin, E. Jan, and J. Flanagan, "Microphone arrays and speaker identification," *IEEE Trans. SAP*, vol. 2, no. 4, pp. 622–629, Oct. 1994.

[4] J. Ortega-Garca and J. Gonzlez-Rodrguez, "Providing single and multi-channel acoustical robustness to speaker identification systems," in *Proc. IEEE ICASSP'97*, Munich, Germany, Apr. 1997, vol. 2, pp.1107–1110.

[5] R.C. Rose, E.M. Hofstetter, D.A. Reynolds, "Integrated models of signal and background with application to speaker identification in noise," *IEEE Trans. SAP*, vol. 2, no. 2, pp. 245–257, Apr. 1994.

[6] T. Matsui, T. Kanno, and S. Furui, "Speaker recognition using HMM

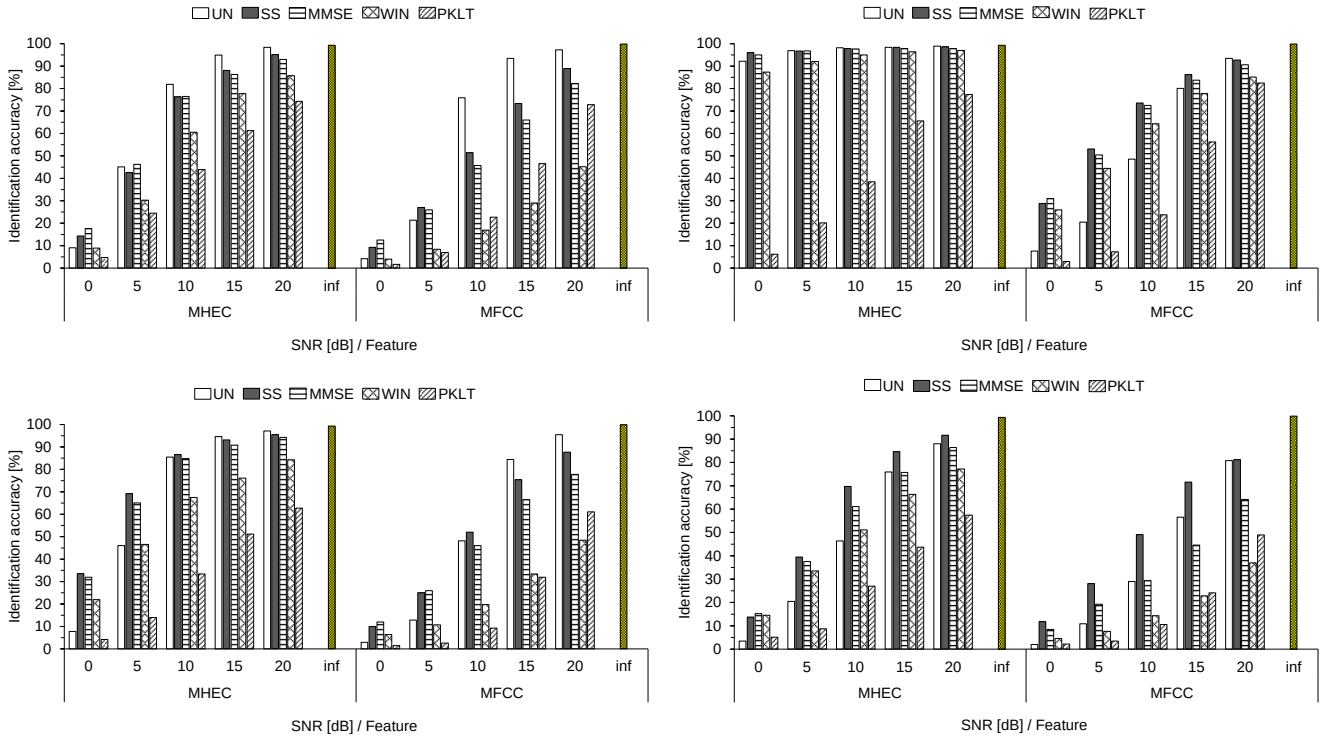

Figure 3: *Performance evaluation of the four speech enhancement front-ends and the MHEC in terms of SID scores under clean and four noisy test conditions including babble (top left), car (top right), factory (bottom left), and white Gaussian (bottom right), at SNRs between* 0–20 *dB.*

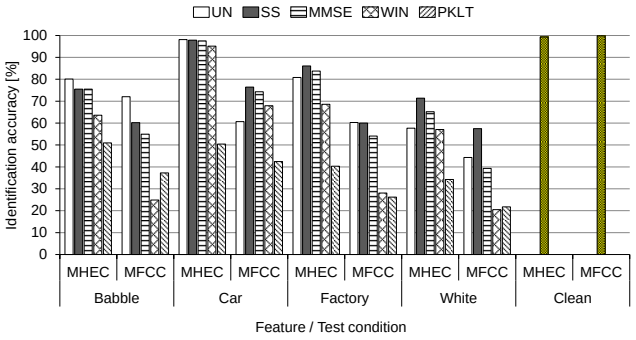

Figure 4: *SID accuracy in clean and four noisy conditions averaged across SNR range* 5–20 *dB, for the un-processed (UN) as well as the processed test materials using the four enhancement front-ends.*

composition in noisy environments," *Compuet. Speech Lang.*, vol. 10, no. 2, pp. 107–116, 1996.

[7] L. Besacier, J. F. Bonastre, and C. Fredouille, "Localization and selection of speaker-specific information with statistical modeling," *Speech Commun.*, vol. 31, pp. 89–106, Jun. 2000.

[8] Y. Shao, S. Srinivasan, D.L. Wang, "Incorporating auditory feature uncertainties in robust speaker identification," in *Proc. IEEE ICASSP'07*, Honolulu, HI, Apr. 2007, vol. IV, pp. 277–280.

[9] J. Ming, T.J. Hazen, J.R. Glass, and D.A. Reynolds, "Robust speaker recognition in noisy conditions," *IEEE Trans. ASLP*, vol. 15, no. 5, pp. 1711–1723, Jul. 2007.

[10] Y. Zhang and W.H. Abdulla "Gammatone auditory filterbank and independent component analysis for speaker identification," in *Proc. INTERSPEECH'06*, Pittsburgh, PA, Sept. 2006, pp. 2098–2101.

[11] T.H. Falk and W.-Y. Chan, "Modulation spectral features for robust far-field speaker identification," *IEEE Trans. ASLP*, vol. 18, no. 1, pp. 90–100, Jan. 2010.

[12] D. A. Reynolds, "Speaker identification and verification using Gaussian mixture speaker models," *Speech Commun.*, vol. 17, pp. 91–108, Aug. 1995.

[13] J.P. Campbell, "Testing with the YOHO CD-ROM voice verification corpus," in *Proc. IEEE ICASSP'95*, Detroit, MI, May 1995, pp. 341–344.

[14] A. Varga and H.J.M. Steeneken, "Assessment for automatic speech recognition: II. NOISEX-92: a database and an experiment to study the effect of additive noise on speech recognition systems," *Speech Commun.*, vol. 12, no. 3, pp. 247–251, Jul. 1993.

[15] M. Berouti, R. Schwartz, J. Makhoul, "Enhancement of speech corrupted by acoustic noise," in *Proc. IEEE ICASSP'79*, Washington, DC, Apr. 1979, pp. 208–211.

[16] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator," *IEEE Trans. ASSP*, vol. ASSP-32, no. 6, pp. 1109–1121, Dec. 1984.

[17] P. Scalart and J. Vieira-Filho, "Speech enhancement based on a priori signal to noise estimation," in *Proc. IEEE ICASSP'96*, Atlanta, GA, May 1996, pp. 629–632.

[18] F. Jabloun and B. Champagne, "Incorporating the human hearing properties in the signal subspace approach for speech enhancement," *IEEE Trans. SAP*, vol. 11, no. 6, pp. 700–708, Nov. 2003.

[19] J. Sohn , N. Kim , and W. Sung, "A statistical model-based voice activity detection," *IEEE Signal Process. Lett.*, vol. 6, no. 1, pp. 1–3, Jan. 1999.

[20] Y. Hu and P.C. Loizou, "Subjective comparison and evaluation of speech enhancement algorithms", *Speech Commun.*, vol. 49, pp. 588–601, 2007.

[21] R.D. Patterson *et al.*, "Complex sounds and auditory images," in *Auditory Physiology and Perception*, Y. Cazals, L. Demany, and K. Horner Eds. Oxford: Pergamon Press, 1992, pp. 429–446.

[22] ITU-T P.56, "Objective measurement of active speech level," ITU-T Recommendation, p. 56, Mar. 1993.

[23] Filtering and Noise Adding Tool (FaNT), available from http://dnt.kr.hs-niederrhein.de/download.html