Robust Speech Enhancement Techniques for ASR in Non-stationary Noise and Dynamic Environments

Gang Liu [1]*, *Dimitrios Dimitriadis*[2], *Enrico Bocchieri*[2]

[1] Center for Robust Speech Systems, University of Texas at Dallas, Richardson, Texas 75083
[2] AT&T Research, 180 Park Ave, Florham Park, New Jersey 07932

gang.liu@utdallas.edu, {ddim, enrico}@research.att.com

Abstract

In the current ASR systems the presence of competing speakers greatly degrades the recognition performance. This phenomenon is getting even more prominent in the case of hands-free, far-field ASR systems like the "Smart-TV" systems, where reverberation and non-stationary noise pose additional challenges. Furthermore, speakers are, most often, not standing still while speaking. To address these issues, we propose a cascaded system that includes Time Differences of Arrival estimation, multi-channel Wiener Filtering, non-negative matrix factorization (NMF), multi-condition training, and robust feature extraction, whereas each of them additively improves the overall performance. The final cascaded system presents an average of 50% and 45% relative improvement in ASR word accuracy for the CHiME 2011(non-stationary noise) and CHiME 2012 (non-stationary noise plus speaker head movement) tasks, respectively.

Index Terms: array signal processing, automatic speech recognition, robustness, acoustic noise, non-negative matrix factorization

1. Introduction

After decades of research, automatic speech recognition (ASR) technology has achieved a performance level that allows commercial deployment. However, most of the current ASR systems fail to perform well in the presence of noise, especially when this noise is non-stationary, e.g., competing speakers. This deterioration in performance is getting even more prominent in the case of hands-free, far-field ASR systems, where overlapping noise sources like the TV audio, traffic noise, and reverberation pose additional challenges [1]. Some of the ongoing research has reported partial success by focusing on either the single channel-based enhancement [2, 3, 15, 16] or the feature extraction side [4, 12, 17]. This study presents a cascaded system that includes time differences of arrival, multi-channel Wiener Filtering, non-negative matrix factorization (NMF), multi-condition training, and robust feature extraction to address these far-field ASR related challenges.

At first, an angular, spectrum-based method is employed to get the spatial information of different audio sources. The generalized cross-correlation with phase transform (GCC-PHAT) method is adopted to estimate the Time Differences of Arrival (TDOA) of all the sources present [9]. The traditional approach either neglects entirely the spatial information or uses a simple channel-averaging method, which yields a sub-optimal performance [2, 3]. However, this work examines how beamforming can benefit the ASR performance, especially when the users are not still. Then, supervised convolutive method [5] is applied to separate speech and filter out some of the competing speech and noise. Herein, the target speech and noise dictionaries are learned and then, applied to the unseen data. Finally, multi-condition training is introduced to cope with the noise mismatches. To further enhance the robustness of the system, two state-of-the-art front-ends, i.e., the ETSI Advanced Front End (AFE) and the Power Normalized Cepstral Coefficients (PNCC), are examined, providing significant improvements in ASR performance.

This paper is structured as follows: First, the relation to prior work is presented in Section 2. The baseline system is outlined in Section 3. The individual "building" blocks and the cascaded system are presented in Section 4, describing how each one of the modules contributes to the overall system performance. Section 5 provides an overview of the speech databases and presents the experimental results of the proposed system with all the intermediate results. The paper is concluded in Section 6.

2. Relation to prior work

This paper presents a cascade of different modules and techniques that are inspired by the systems presented in CHiME2011 challenge [1]. The novelty of this paper is outlined in the successful combination of different, multi-discipline algorithms such as the feature extraction scheme combined with multi-channel Wiener filtering. Prior work, as in [3, 10, 12], presents the original algorithms but herein, the proposed system gracefully combines them, outperforming systems as those presented in the CHiME challenges. Additionally, this work also investigates other robust feature extraction schemes, i.e., PNCC and AFE, and validates their performance on the challenging noisy data. Finally, the spatial information compensating for the speakers' head movements was seldom examined before.

3. Baseline system

The baseline system is trained on 39-dimensional MFCC features, i.e. 12 MFCC and a log-energy coefficient plus their delta and acceleration coefficients and Cepstral Mean Normalization. The words are modeled by whole-word left-to-right HMMs, with no skips over states and 7 Gaussian mixtures per state with diagonal covariance matrices. The state number for each word model is approximately 2 states per phoneme. The baseline system is based on the HTK [8] and provided by [6] to assist comparison. Note that the back-end system remains the same in all experiments presented herein.

4. Proposed cascaded system

The proposed cascaded system is illustrated in Fig. 1 and the "building" modules are presented below. Some separation examples are available from the demo web page [18].

*This work was performed while interning at AT&T Research

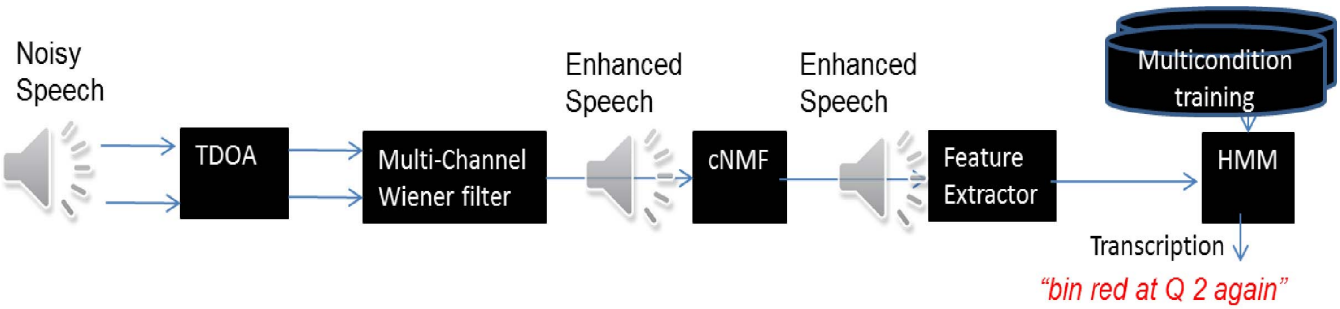

Figure 1: *Flowchart of the proposed system: speech separation + noise suppression + robust feature + Multicondition training. The input is a two-channel audio file. After the "Multi-Channel Wiener filter", the audio file becomes mono-channel. For the ground-truth transcription "bin red at Q 2 again", only the 4th and the 5th words need to be recognized.*

4.1. Time Difference of Arrival (TDOA) and Multi-channel Wiener filtering (MWF)

One of the advantages of microphone arrays is their spatial discrimination of the acoustic sources. The problem of source localization/separation is often addressed by the TDOA estimation. In this study, we focus on the TDOA estimation of two or more sources for the given pair of sensors. The estimation is based on the Short Time Fourier Transform (STFT) of the mixed signal $x(t,f)$:

$$x(t,f)=\sum_{n=1}^{N} d(f,\tau_n)s_n(t,f)+b(t,f)\,, \tag{1}$$

where $d(f,\tau_n)=[1,e^{-2\pi i f\tau_n}]^T$ is the steering vector associated with the n^{th} source τ_n, $b(t,f)$ models the residual noise, $x(t,f)$ and $s_n(t,f)$ are, respectively, the STFTs of the observed signals and the n^{th} source signal, where $t=1,...,T$, $f=1,...,F$, and $n=1,...,N$ are the time frame, frequency bin, and source indices, respectively.

The "local" [1] angular spectrum $\phi(t,f,\tau)$ is computed for every *t-f* bin and for all possible values of τ lying on a uniform grid space. Artifacts are introduced due to either the spatial aliasing occurring especially at the high frequencies or the irrelevant information introduced when the desired sound source is inactive. Thus, a maximization operation is adopted on the angular spectrum based on the Generalized Cross-Correlation with PHase Angular spectrum Transform (GCC-PHAT) [9], to make the process more robust.

$$\phi^{\max}(\tau)=\arg\max_{t}\sum_{f=1}^{F}\phi(t,f,\tau) \tag{2}$$

where $\phi(t,f,\tau)$ is the "local" angular spectrum.

After time-aligning the observed mixed signals based on their TDOA estimate, a Multi-channel Wiener Filtering (MWF) is used to suppress most of the stationary noise [14]. The Wiener filter W_n is computed as:

$$W_n(t,f)=R_{c_n}(t,f)R_X^{-1}(t,f) \tag{3}$$

where R_X and R_{c_n} is the covariance of the spatial image[2] of the mixture and the n^{th} source, respectively. Henceforth, this procedure is noted as TDOA + MWF and it is shown to perform similarly to a beamformer [9]. This step provides significant performance improvements in Section 5.

4.2. Convolutive NMF (cNMF)

MWF suppresses the stationary noise but the processed audio signals may still contain a residual of non-stationary noise e.g., non-target speech, music, etc. To deal with that, cNMF is introduced to further enhance the speech quality [10]. Assuming the speech is corrupted by additive noise:

$$V=V^{(s)}+V^{(n)}\,, \tag{4}$$

where V , $V^{(s)}$ and $V^{(n)}$ are the non-negative matrices representing the magnitude spectrogram of the audio signal, the targeted speech, and the noise (or non-target speech) signal, respectively, following the notation in [10]. In practice, V is estimated as:

$$V\approx\hat{V}^{(s)}+\hat{V}^{(n)}=\sum_{p=0}^{P-1}W^{(s)}(p)\overrightarrow{H^{(s)}}+\sum_{p=0}^{P-1}W^{(n)}(p)\overrightarrow{H^{(n)}} \tag{5}$$

where $\overrightarrow{H^{(s)}}$ is a 'shifted' version of non-negative matrix $H^{(s)}$ where the entries are shifted *p* turns to the right, filling with zeros from the left. The speech and noise dictionaries, $W^{(s)}(p)$ and $W^{(n)}(p)$ are estimated from the clean targeted and non-targeted speech training data, respectively. The openBliSSART toolkit [10] is used for the cNMF process.

In more detail, the cNMF-based speech enhancement can be summarized as:

1) Train two dictionaries offline, one for speech (speaker dependent dictionary) and one for noise.
2) Calculate the magnitude of the STFT of the mixed speech signal.
3) Separate the magnitude into two parts via cNMF, each of which can be sparsely represented by either the speech or the noise dictionaries.
4) Reconstruct speech from the part corresponding to the speech dictionary.

[1] The term 'local' emphasizes the fact that the angular spectrum is estimated on a local grid searching basis. It is an approximation method.

[2] 'Spatial image' is the contribution of a sound source to all mixture channels.

4.3. Feature Extraction

After the noise suppression process, there are still some non-target noise residuals. Therefore, two state-of-the-art feature extraction schemes are investigated.

The first feature extraction scheme is based on the ETSI Advanced Front End (AFE) [11]. For AFE, the noise reduction scheme is based on a two-step Wiener filtering (WF): First, a voice activity detector is applied to label the speech frames. Based on this speech/noise decision, a WF is estimated at the *mel* filter-bank energy domain for each t-f bin. The WF is then applied to the input waveform and the denoised time-domain signal is reconstructed. The entire process is repeated twice and then, MFCCs are extracted.

The second front-end examined is the Power Normalized Cepstral Coefficients (PNCC) [12]. The difference with MFCCs is the introduction of the power-law nonlinearity that replaces the traditional log-nonlinearity, and a noise-suppression algorithm based on asymmetric filtering that suppresses the background noise. This asymmetric noise suppression scheme is based on the observation that the speech energy in every channel usually changes faster than the background noise energy in the same channel. As shown in [12], these features outperform the MFCCs in reverberated speech.

4.4. Multi-condition Training

Finally, multi-condition training (MCT) approximates the test set by creating training data with matched noise conditions, e.g., adding noise of various signal-to-noise ratio (SNR) levels [13]. The MCT training data are created by mixing clean reverberated data with isolated background noise sequences for six different SNR levels. The outcome of this process matches closer to the development/test sets [6].

5. Experiment and results

The overall performance of the proposed system is examined in the two released 2011 and 2012 'CHiME' speaker-dependent, small vocabulary ASR tasks.

5.1. The 2011 and 2012 CHiME Corpus

The PASCAL 2011 'CHiME' Speech Separation and Recognition Challenge [6] is designed to address some of the problems occurring in real-world noisy conditions. The data from this challenge are based on the GRID corpus [7], where 34 speakers read simple command sentences. These sentences are in the form *'verb-color-preposition-letter-digit-adverb'*. There are 25 different *'letter'*-class words and 10 different *'digit'*-class words. Other classes have a four-word option each. In the CHiME recognition task, the ASR performance is measured in the percentage of correctly recognized 'letter' and 'digit' keywords (termed as correct word accuracy here). The CHiME data simulate the scenario where sentences are spoken in a noisy living room. The original, clean speech utterances are convolved with the room impulse response, and then mixed with random noise signals to target SNR levels of 9, 6, 3, 0, -3 and -6 dB. For training, 500 reverberated utterances per speaker (no noise), and six hours of background noise are used. The development and test sets consist of 600 multi-speaker utterances at each one of the SNR levels. All utterances are given both in end-pointed format and noise as embedded signals. All data are stereo recorded in 16 kHz sampling frequency.

The main difference between the CHiME 2011 data and the CHiME 2012 data is that the target speaker is now allowed to make small movements within a square zone of +/- 0.1m around the center position [1].

5.2. Results

Herein, we present the cascaded system consisting of the aforementioned sub-components. The contribution of each one of the sub-components is detailed in Table 1. First of all, it is observed that the proposed systems offer consistent, additive improvements on both the development and the test sets, when compared with the baseline system performance. Without any loss of generality, we will focus on, and provide more details about the test set of CHiME 2011, although similar general conclusions apply to the rest of the experiments, as well. The first two rows in Table 1 use just channel-averages yielding sub-optimal results. As shown in Fig. 2, NMF provides the largest boost, due to the suppression of the non-stationary interfering signals. The second largest improvement stems from the TDOA+MWF module. Additional improvement comes from the MCT process. Although MCT dramatically increases the training time, it can be done off-line and so it is still practical. The last row of Table 1 details the relative gain of the cascaded against the baseline system. It is observed that although the proposed system always improves the ASR performance, it favors the most adverse environment (i.e., -6dB), presenting a 138% relative improvement in terms of word accuracy. Finally, it is noted that PNCC and AFE can offer significant better performance than MFCC.

Similarly, the proposed cascaded system is also applied on the CHiME 2012 data with results detailed in Table 2. Compared to the baseline system, the proposed system provides 42% and 45% relative improvement for the development and test scenarios. However, the proposed system doesn't offer similar improvements as those reported in CHiME 2011 due to the movement of the target speaker. The TDOA module provides rough-grain estimates in the sentence level; therefore some of the head movements remain unattended. Head tracking algorithms presenting finer time-resolution will be examined next. Herein, it is assumed that the speaker moves his head only between sentences. This assumption seems to be reasonable when comparing Table 1 and 2. The difference between the two test sets of CHiME 2011 and 2012 is only 0.8%, which is almost negligible.

6. Conclusion

In this paper, we propose a cascaded system for speech recognition dealing with non-stationary noise in reverberated environments. It efficiently copes with speaker movements. The proposed system offers an average of 50% and 45% in relative improvements for the above mentioned two scenarios, respectively. Although most of the reported improvements

Table 1. *CHiME2011: Comparison of Correct Word Accuracies in % for the development (left) and test set (right). (TDOA: Time of Difference of Arrival; MWF: Multi-channel Wiener Filtering; NMF: Non-negative Matrix Factorization; MCT: Multi-condition Training; '+' means cascading; Relative gains are computed against baseline system).*

System Setup	Correct Word Accuracy (%) on **Development Set**							Correct Word Accuracy (%) on **Test Set**						
	-6dB	-3dB	0 dB	3 dB	6 dB	9 dB	Avg.	-6dB	-3 dB	0 dB	3 dB	6 dB	9 dB	Avg.
CHiME 2011 Baseline	31.1	36.8	49.1	64.0	73.8	83.1	56.3	30.3	35.4	49.5	62.9	75	82.4	55.9
Baseline + Wiener Filtering	31.4	42.1	54.2	65.2	76.9	83.3	58.9	31.3	40.5	53.2	65.4	76.4	83.3	58.4
TDOA+MWF	44.8	49.8	60.2	69.9	78.0	82.5	64.2	44.1	48.1	61.7	70.6	79.9	83.5	64.6
TDOA+MWF+NMF	60.2	67.9	73.6	79.0	83.6	84.8	74.8	64.7	68.2	75.4	81.2	83.7	86.3	76.6
TDOA+MWF+ NMF+MCT (MFCC)	67.9	71.6	77.8	82.4	84.8	86.2	78.5	69.8	74.0	79.2	84.0	87.2	90.2	80.7
TDOA+MWF+ NMF+MCT(PNCC)	70.3	74.2	79.4	83.8	**87.0**	87.6	80.4	74.6	78.8	**83.5**	84.7	88.2	89.8	83.2
TDOA+MWF+ NMF+MCT (AFE)	**73.8**	**77.3**	**82.4**	**85.5**	86.4	**88.9**	**82.4**	**75.8**	**79.2**	83.4	**86.8**	**88.7**	**90.8**	**84.1**
Relative Gain (%)	138	110	68	34	17	7	46	150	124	68	38	18	10	50

Table 2. *CHiME2012: Comparison of Correct Word Accuracies in % for the development (left) and test set (right).*

System Setup	Correct Word Accuracy (%) on **Development Set**							Correct Word Accuracy (%) on **Test Set**						
	-6dB	-3 dB	0 dB	3 dB	6 dB	9 dB	Avg.	-6dB	-3 dB	0 dB	3 dB	6 dB	9 dB	Avg.
CHiME 2012 Baseline	32.1	36.3	50.3	64.0	75.1	83.5	56.9	32.2	38.3	52.1	62.7	76.1	83.8	57.5
TDOA+MWF+NMF+MCT (PNCC)	**69.8**	**76.8**	81.1	84.6	**86.8**	**88.7**	**81.3**	72.8	76.8	82.3	**86.3**	88.1	89.1	82.3
TDOA+MWF+NMF+MCT (AFE)	**69.8**	75.8	**81.2**	84.2	86.1	87.3	80.7	**73.7**	**78.1**	**83.8**	85.8	**88.8**	**89.9**	**83.3**

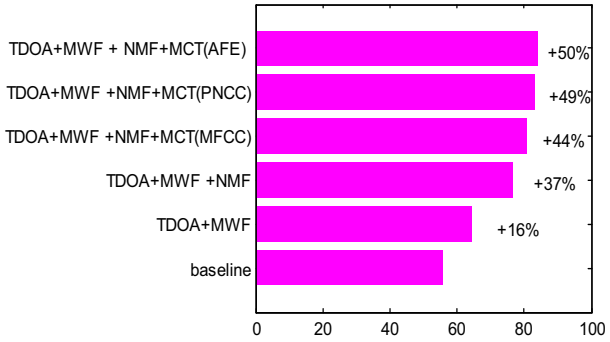

Figure 2: *Relative improvements against baseline by introducing a single new component to the system. Results reported on the CHiME2011 test set (based on average word accuracy).*

come from the signal processing domain, further improvements can be derived by introducing robust acoustic modeling methods such as model adaptation, e.g., MLLR.

As next steps of this work, the spatial information and the speaker movements will be further exploited aligning the system with a real-life working scenario. Due to the generic property of the framework, it can be further validated in other noisy speech challenge, such as emotion identification [19-20], language identification [21-23], speaker identification [24-30], and speaker counting [31-32].

Part of the core code to produce the results in this paper may be found at *http://www.utd.edu/~gang.liu/demo_ATT_CHIME_denoise.htm.*

7. Acknowledgements

The authors would like to thank Felix Weninger for his help in the NMF experiments and Roland Maas for sharing the multi-condition training scripts. Thank anonymous reviewers for their valuable comments.

8. References

[1] J. Barker, E. Vincent, N. Ma, H. Christensen, and P. Green, "The PASCAL CHiME Speech Separation and Recognition Challenge," Computer Speech and Language, vol. 27, no. 3, pp. 621–633, 2013.

[2] J. F. Gemmeke, T. Virtanen, and A. Hurmalainen. "Exemplar-Based Speech Enhancement and its Application to Noise-Robust Automatic Speech Recognition". Proc. 1st Int. Workshop on Machine Listening in Multisource Environments (CHiME), pp. 53–57, 2011.

[3] F. Weninger, J. Geiger, M. Wöllmer, B. Schuller, G. Rigoll, "The Munich 2011 CHiME challenge contribution: NMF-BLSTM speech enhancement and recognition for reverberated multisource environments", Proc. 1st Int. Workshop on Machine Listening in Multisource Environments (CHiME), pp. 24–29, 2011.

[4] F. Nesta, M. Matassoni, "Robust automatic speech recognition through on-line semi blind source extraction", Proc. 1st Int. Workshop on Machine Listening in Multisource Environments (CHiME), pp. 18–23, 2011.

[5] P. Smaragdis, "Convolutive speech bases and their application to supervised speech separation," IEEE Trans. on Audio, Speech and Language Processing, Vol. 15, No 1, pp. 1–14, 2007.

[6] J. Barker, H. Christensen, N. Ma, P. Green, and E. Vincent. "The PASCAL CHiME speech separation and recognition challenge", 2011. [Online]. Available: http://www.dcs.shef.ac.uk/spandh/chime/challenge.html.

[7] M. Cooke, J. Barker, S. Cunningham, X. Shao, "An audio-visual corpus for speech perception and automatic speech recognition". Journal of the Acoustical Society of America, Vol. 120, pp. 2421–2424, 2006.

[8] S. Young, G. Evermann, M. J. F. Gales, T. Hain, D. Kershaw, X. Liu, G. Moore, J. J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland, The HTK Book (for version 3.4), Cambridge University Engineering Department, 2009.
[9] C. Blandin, A. Ozerov and E. Vincent, "Multi-source TDOA estimation in reverberant audio using angular spectra and clustering", Signal Processing, Elsevier, 2011.
[10] B. Schuller, A. Lehmann, F. Weninger, F. Eyben, G. Rigoll, "Blind Enhancement of the Rhythmic and Harmonic Sections by NMF: Does it help?", in Proc. NAG/DAGA, pp. 361-364, 2009.
[11] Speech Processing, Transmission and Quality Aspects (STQ); Distributed Speech Recognition; Advanced Front-end Feature Extraction Algorithm; Compression Algorithms, European Telecommunications Standards Institute ES 202 050, Rev. 1.1.5, Jan. 2007.
[12] C. Kim and R. M. Stern. "Power-normalized coefficients (PNCC) for robust speech recognition", in IEEE. Conf. Acoust, Speech, and Signal Processing, pp.4101-4104, March, 2012.
[13] J. Ming, T.J. Hazen, J.R. Glass, and D.A. Reynolds. "Robust speaker recognition in noisy conditions". IEEE Trans. Audio, Speech and Language Processing, Vol. 15, No 5, pp.1711 – 1723, July 2007.
[14] N. Q. K. Duong, E. Vincent, and R. Gribonval, "Under-determined reverberant audio source separation using a full-rank spatial covariance model", IEEE Trans. on Audio, Speech, and Language Processing, vol. 18, no. 7, pp. 1830–1840, 2010.
[15] J. Wu, J. Droppo, L. Deng, and A. Acero, "A noise-robust ASR front-end using wiener filter constructed from MMSE estimation of clean speech and noise", in Proc. ASRU, pp. 321-326, 2003.
[16] M. Cooke, J. R. Hershey, and S. J. Rennie, "Monaural speech separation and recognition challenge". Computer Speech & Language, vol. 24, no. 1, pp.1-15. 2010.
[17] U. H. 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.
[18] "Demo of speech de-noise in non-stationary noise and dynamic environments" [Online]. Available: http://www.utd.edu/~gang.liu/demo_ATT_CHIME_denoise.htm
[19] G. Liu, Y. Lei, and J.H.L. Hansen, "A Novel Feature Extraction Strategy for Multi-stream Robust Emotion Identification", in Proc. INTERSPEECH. Makuhari Messe, Japan, pp.482-485, 2010.
[20] 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, pp.3285-3288, 2011.
[21] G. Liu and J.H.L. Hansen, "A systematic strategy for robust automatic dialect identification", in Proc. EUSIPCO, Barcelona, Spain, pp.2138-2141, 2011.
[22] G. Liu, Y. Lei, and J.H.L. Hansen, "Dialect Identification: Impact of difference between Read versus spontaneous speech", in Proc. EUSIPCO, Aalborg, Denmark, pp.2003-2006, 2010.
[23] G. Liu, C. Zhang, and J.H.L. Hansen, "A Linguistic Data Acquisition Front-End for Language Recognition Evaluation", in Proc. Odyssey, Singapore, pp. 224-228, 25-28 June 2012.
[24] G. Liu, Y. Lei, and J.H.L. Hansen, "Robust feature front-end for speaker identification", in Proc. ICASSP, Kyoto, Japan, pp.4233-4236, 2012.
[25] 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, pp. 7755-7759, 2013.
[26] G. Liu, J.-W. Suh, and J.H.L. Hansen, "A fast speaker verification with universal background support data selection", in Proc. ICASSP, Kyoto, Japan, pp.4793-4796, 2012.
[27] R. Saeidi, et al., "I4U submission to NIST SRE 2012: A large-scale collaborative effort for noise-robust speaker verification", accepted to INTERSPEECH, Lyon, France, 25-29 Aug., 2013.
[28] J.-W. Suh, S. O. Sadjadi, G. Liu, T. Hasan, K. W. Godin, and J. H.L. Hansen, "Exploring hilbert envelope based acoustic features in i-vector speaker verification using HT-PLDA," in Proc. NIST Speaker Recognition Evaluation Workshop, 2011.
[29] V. Hautamaki, et al., "Automatic regularization of cross-entropy cost for speaker recognition fusion", accepted to INTERSPEECH, Lyon, France, 25-29 Aug., 2013.
[30] 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, pp. 6783-6787, 2013.
[31] C. Xu, S. Li, G. Liu, Y. Zhang, E. Miluzzo, Y. Chen, J. Li, and B. Firner, "Crowd++: Unsupervised Speaker Count with Smartphones," accepted in the 2013 ACM International Joint Conference on Pervasive and Ubiquitous Computing (ACM UbiComp), Zurich, Switzerland, September 9-12, 2013.
[32] C. Xu, S. Li, G. Liu, Y. Zhang, E. Miluzzo, Y. Chen, J. Li, and B. Firner, "Unsupervised Speaker Counting with Smartphones," accepted in the 2013 ACM International Joint Conference on Pervasive and Ubiquitous Computing (ACM UbiComp), Poster Session, Zurich, Switzerland, September 9-12, 2013.