IMPROVING MICROPHONE ARRAY SPEECH RECOGNITION WITH COCHLEAR IMPLANT-LIKE SPECTRALLY REDUCED SPEECH

Cong-Thanh Do Mohammad J. Taghizadeh
Philip N. Garner

Idiap-RR-40-2011

DECEMBER 2011

Centre du Parc, Rue Marconi 19, P.O. Box 592, CH - 1920 Martigny
T +41 27 721 77 11 F +41 27 721 77 12 info@idiap.ch www.idiap.ch

IMPROVING MICROPHONE ARRAY SPEECH RECOGNITION WITH COCHLEAR IMPLANT-LIKE SPECTRALLY REDUCED SPEECH

Cong-Thanh Do[1], Mohammad J. Taghizadeh[1,2] and Philip N. Garner[1]

September 28, 2011

Abstract

Cochlear implant-like spectrally reduced speech (SRS) has previously been shown to afford robustness to additive noise. In this paper, it is evaluated in the context of microphone array based automatic speech recognition (ASR). It is compared to and combined with post-filter and cepstral normalisation techniques. When there is no overlapping speech, the combination of cepstral normalization and the SRS-based ASR framework gives a performance comparable with the best obtained with a non-SRS baseline system, using maximum a posteriori (MAP) adaptation, either on microphone array signal or lapel microphone signal. When there is overlapping speech from competing speakers, the same combination gives significantly better word error rates compared to the best ones obtained with the previously published baseline system. Experiments are performed with the MONC database and HTK toolkit.

Keywords: Cochlear implant, Microphone array, Noise robust ASR, Overlapping speech, Spectrally reduced speech

1 Introduction

Speech recognition in meetings presents an important application domain for speech recognition technologies [1]. However, this is a difficult recognition task. Apart from ambient background noise and reverberation, a major source of noise in meetings is overlapping speech from other participants. These overlapped speech segments cause problems for speech recognition and induce significant increase in word error rate.

The use of microphone arrays can help in improving significantly speech recognition performance in meetings, compared to the performance of lapel microphone speech recognition, when there is overlapping speech [1]. The enhancement-based approach [2] is the most common method of performing speech recognition using a microphone array [3]. In this approach, either a fixed or adaptive beamforming algorithm is applied to the multi-channel captured audio. Indeed, noise in the received signal can be significantly reduced thanks to these beamforming algorithms. However, these algorithms cannot remove noise entirely from the received signals in any realistic environment. Consequently, the single-channel output signal, generated by a beamforming stage, is consecutively processed with a post-filter. The post-filtering can be performed using conventional single channel speech enhancement algorithms, e.g. Wiener filtering or spectral subtraction algorithms [4], or can be built based on information extracted from all array channels. The single-channel output signal from the post-filter is then passed to the ASR system for feature extraction and decoding. This enhancement-based approach is widely used since it is simple and gives comparable recognition performance compared to other approaches, e.g., the multi-stream approach, which are more computationally costly [3].

In this paper, we investigate the possibility of improving the performance of enhancement-based microphone array speech recognition using cochlear implant-like spectrally reduced speech (SRS), which is the acoustic simulation of a cochlear implant [5]. Indeed, a novel framework for noise robust ASR has been recently introduced based on the SRS [6]. In this framework, the SRS signals, synthesized from original clean speech signals, are used for training; the SRS signals, synthesized from noisy speech signals, are used for testing. It has also been suggested that the implementation of other noise robust techniques, e.g., speech enhancement, on this SRS-based framework could further improve noise robustness [6]. We thus implement some standard noise robust techniques on this framework and observe the ASR performance. The results obtained with the SRS-based framework are compared with those obtained with the baseline framework using standard noise robust techniques. Experiments are performed on the multichannel overlapping numbers corpus (MONC) [7].

The paper is organized as follows. Section 2 describes the SRS synthesis algorithms. In section 3, the microphone array processing is introduced. Section 4 and section 5 present the experimental setup and the recognition results, respectively. Finally, section 6 concludes the paper.

2 SRS Synthesis Algorithm

In the SRS technique [6], a speech signal $s(t)$ is first decomposed into N subband signals, $s_i(t), i = 1, \ldots, N$, by using a perceptually-motivated analysis filterbank consisting of N bandpass filters. The aim of the analysis filterbank is to simulate the motion of the basilar membrane [8]. In this respect, the filterbank consists of nonuniform bandwidth bandpass filters that are linearly spaced on the Bark scale. In this paper, each bandpass filter in the filterbank is a second-order elliptic bandpass filter having a minimum stopband attenuation of 50 dB and a 2 dB peak-to-peak ripple in the passband. The lower, upper, and central frequencies of the bandpass filters are calculated as in [9]. An example of analysis filterbank is given in Fig. 1

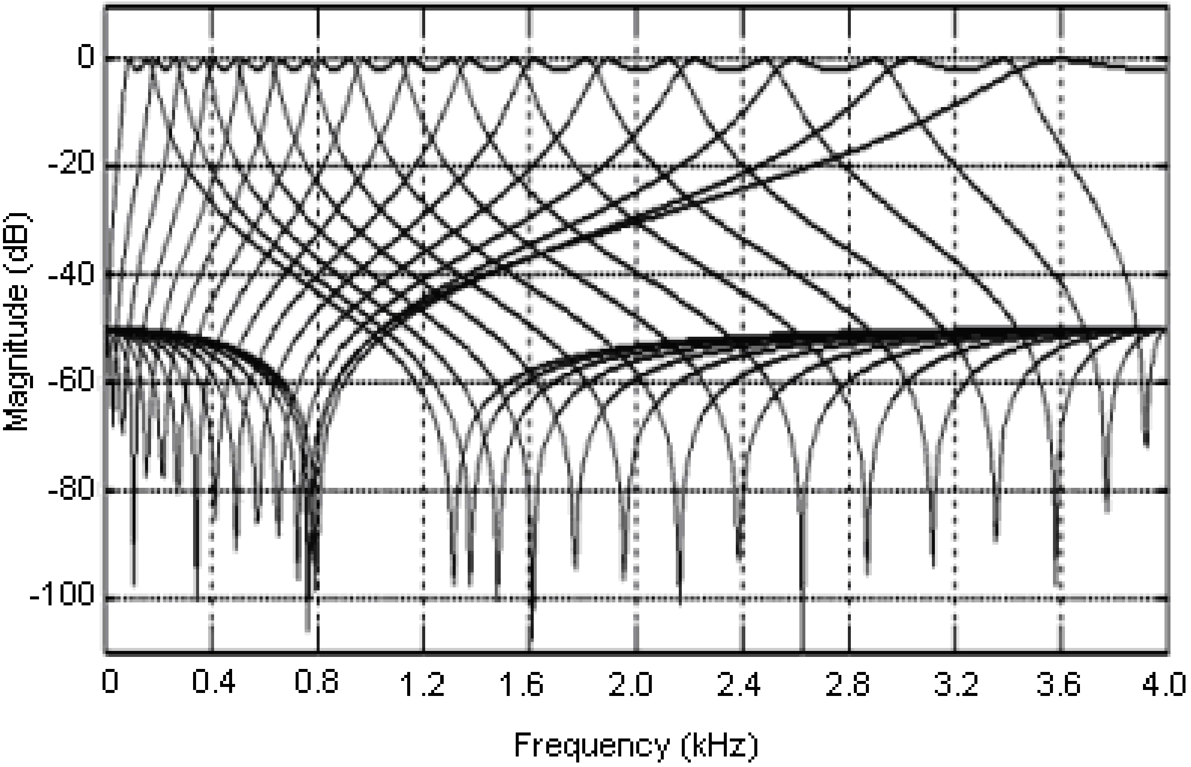

Figure 1: Frequency response of an analysis filterbank consisting of 16 second-order elliptic bandpass filters used for speech signal decomposition. The speech signal is sampled at 8 kHz.

The amplitude modulations (AMs), $m_i(t)$, of the subband signals, $s_i(t), i = 1, \ldots, N$, are then extracted by, first, full-wave rectification of the outputs of the bandpass filters and, subsequently, lowpass filtering of the resulting signals. The sampling rate of the AM is kept at the same value as that of the subband signal (8 kHz). In this work, the AM filter is a fourth-order elliptic lowpass filter with 2 dB of peak-to-peak ripple and a minimum stop-band attenuation of 50 dB. The subband AM, $m_i(t)$, is then used to modulate a sinusoid whose frequency, f_{ci}, equal to the central frequency of the corresponding analysis bandpass filter of that subband. Afterwards, the subband modulated signal is spectrally limited (i.e., is filtered again) by the same bandpass filter used for the original analysis subband [5]. Finally, all the subband spectrally limited signals are summed to synthesize the SRS. The SRS, $\widehat{s}(t)$, can be expressed as follows, and the SRS synthesis algorithm is summarized in Fig. 2:

$$\widehat{s}(t) = \sum_{i=1}^{N} m_i(t) \cos(2\pi f_{ci} t) \tag{1}$$

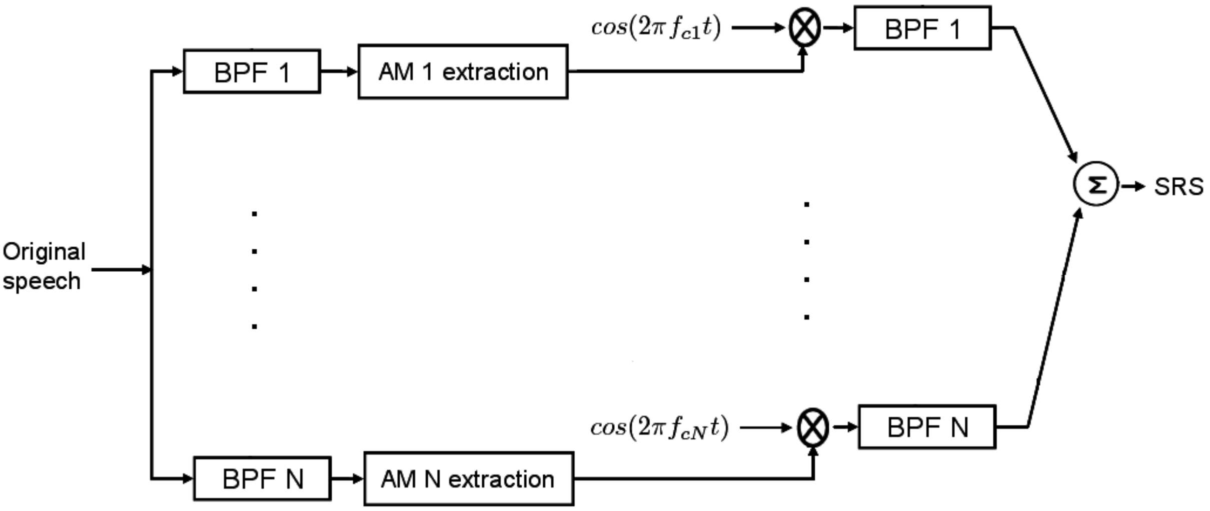

Figure 2: SRS synthesis algorithm [5]. The original speech signal is decomposed into several subband speech signals. The subband temporal envelopes are extracted from the subband speech signals and are then used to synthesize the SRS. BPF means bandpass filter.

The use of SRS signals, which contains basic and ASR relevant information transfered by the subband temporal envelopes [10], improves ASR noise robustness since the use of such basic information in ASR helps in reducing variability from the speech signal, especially when the speech is contaminated by environmental noise [6].

3 Microphone Array Processing

Microphone arrays are designed for high quality acquisition of distant speech, relying on beamforming or spatial filtering. The directionally discriminative time-space filtering of multi-channel acquisition results in suppression of interference sources, thus improving the signal to noise ratio (SNR). Beamforming filters are designed based on requirements of the application. The optimal beamforming for maximizing the array-gain is known as the superdirective beamformer. The array-gain is defined as the SNR improvement of the beamformer output with respect to the single channel.

Figure 3 illustrates the beam-pattern of a superdirective beamformer at frequencies 250, 500, 1000 and 2246 Hz for the microphone array set-up of our recordings [7]. 2246 Hz is the frequency corresponding to the microphone separation being a half wavelength. The speaker is located at azimuth and elevation 135° and 25° respectively with respect to the center of the array. As the figure shows, the beam-pattern is adjusted towards the speaker, and it is kept the same for all scenarios. The average SNR of the recordings is 9 dB. The dominant noise has diffuse characteristics [11] so we use a McCowan post-filter to achieve a higher accuracy using the superdirective beamformer. The filter assumes that we know the noise field coherence function, so a more accurate estimated signal power spectral density is possible.

4 Experimental Setup

4.1 Database

We use the multichannel overlapping numbers corpus (MONC) database [1] for the experiments in this paper. The utterances in the numbers corpus (30-word vocabulary) include isolated digit strings, continuous digit strings, and ordinal/cardinal numbers, collected over telephone lines. For acquiring the MONC database, the utterances of the OGI numbers corpus were played back on one or more loudspeakers,

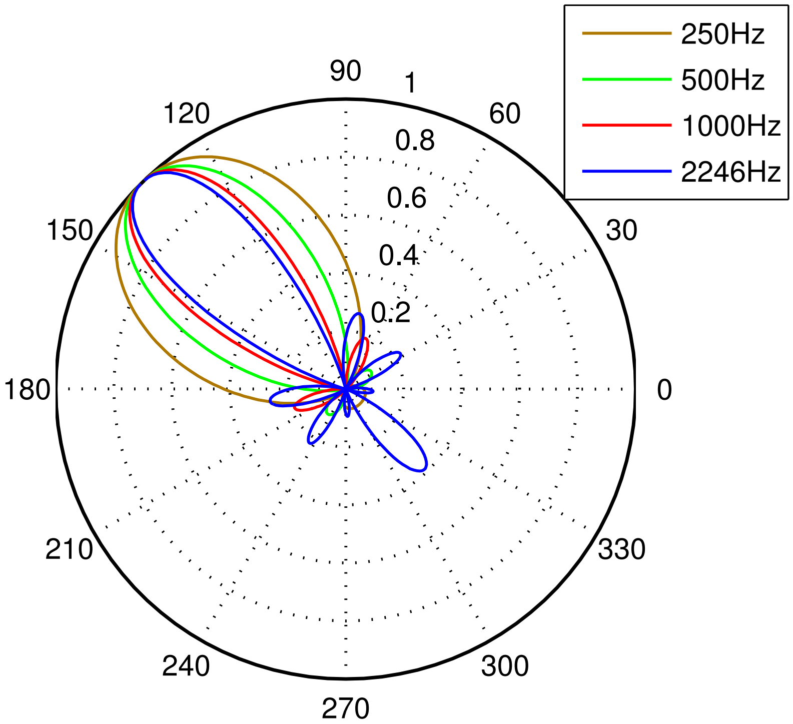

Figure 3: Beam-patterns for superdirective beamformer with circular microphone array.

and the resulting sound field was recorded with lapel microphones, a single tabletop microphone, and a tabletop microphone array. The recordings were made in a moderately reverberant 8.2 m × 3.6 m × 2.4 m rectangular room. Background noise was made mainly by the PC power supply fan. The loudspeakers were positioned around a circular meeting room table to simulate the presence of 3 competing speakers in the meeting room. The angular spacing between them was 90° and the distance from table surface to centre of main speaker element was 35 cm. Lapel microphones were attached to t-shirts hanging below each loudspeaker. The microphone array includes 8 microphones, which were distributed in a 20 cm diameter circle placed in the centre of the table. An additional microphone was placed at the centre of the array. A graphical description of the room arrangement can be found in [7].

As demonstrated by Moore and McCowan [1], when no overlapping speech was present, the microphone array output recognition performance was equivalent to that of the lapel microphone. Otherwise, in the presence of overlapping speech, the microphone array successfully enhanced the desired speech, and gave the better recognition performance compared to that obtained with the lapel microphone. In this paper, we applied various simple noise robust techniques to the output of the microphone array in order to improve ASR performance. We use thus the clean training set for training ASR systems. In addition, we use the outputs of the microphone array corresponding to three recording scenarios, including S_1 (no overlapping speech), S_{12} (one competing speaker), and S_{123} (two competing speakers) as input testing speech of the ASR systems.

4.2 ASR Systems Training

A baseline ASR system was trained in the spirit of that of Moore and McCowan [1], using the HTK toolkit [12], on the clean training set of the original numbers corpus. The system consists of acoustic models which are tied-state triphone hidden Markov models (HMMs). The triphone HMMs are standard with 3 emitting states per triphone and 12 Gaussian mixtures per state. The system uses 39-dimensional speech feature vectors which consist of 13 MFCC coefficients (including the 0th cepstral coefficient) along with their delta and acceleration coefficients. This baseline system gave a word error rate (WER) of 6.45% using the clean test set from the original numbers corpus. Other baseline systems are those trained on original clean speech and tested on the same lapel microphone signals, using maximum a posteriori (MAP) adaptation, as well as on microphone array signals, using MAP adaptation [1].

In the framework of noise robust ASR using cochlear implant-like spectrally reduced speech (SRS), the SRS signals are synthesized from clean training speech signals; then these SRS signals are used to train the ASR system. In testing, the SRS signals are synthesized from the normal test speech signals. In the ASR framework that does not use SRS, original clean speech signals are used to train the ASR system for recognizing normal test speech signals. The initial input testing speech of the two ASR systems are the same: single-channel audio stream after the post-filtering output of the microphone array. This audio stream will be recognized directly with the normal ASR system. Otherwise, SRS signals will be synthesized from this input audio stream and recognized with the SRS-based ASR system. The SRS synthesis needs two important parameters: the number of frequency subbands and the subband temporal envelopes bandwidth. Following earlier results [6], we synthesize 16-subband SRS with 50 Hz subband temporal envelope bandwidth. The experimental protocols are summarized in Fig. 4.

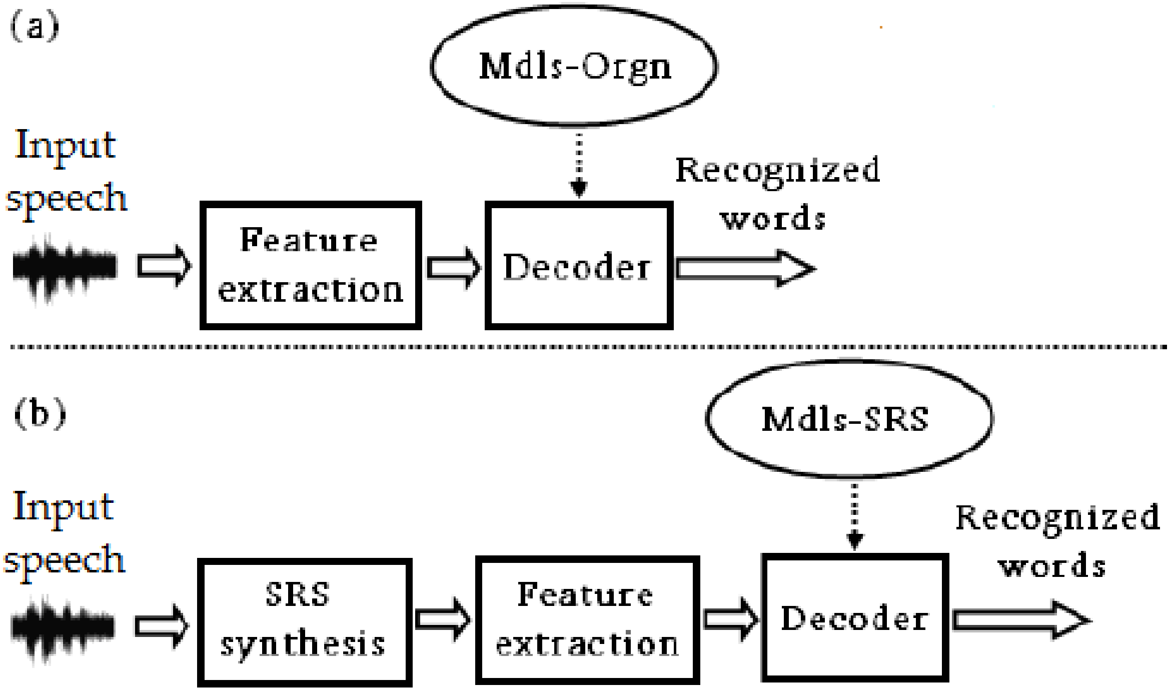

Figure 4: Two experimental frameworks using ASR systems trained on original clean speech (Fig. 4(a)) and on SRS signals (Fig. 4(b)), respectively. In Fig. 4(a), Mdls-Orgn denotes the models (acoustic and language models, dictionary) obtained from training on original clean speech signals. Similarly, in Fig. 4(b), Mdls-SRS denotes the models obtained from training on SRS signals synthesized from original clean speech signals.

5 Recognition Results

We implement cepstral normalization techniques on the SRS-based framework. Cepstral mean normalization (CMN) and cepstral variance normalization (CVN) are in turn known to perform as well as or better than standard noise robust ASR techniques such as spectral subtraction [13]. CMN and CVN are implemented on both ASR framework, whether based on SRS or not. The results obtained with the two frameworks are compared.

Tab. 2 shows the recognition results, in terms of WERs, obtained with two ASR frameworks using noise robust techniques above. We use the WER computed by the baseline ASR system, trained on original clean speech, tested on the same microphone array and lapel microphone signals, using MAP adaptation as noise robust technique, as the references. These WERs are extracted from [1] and are displayed in Tab. 1.

We can remark that, in the S_1 scenario (no overlapping speech), the implementation of CMN alone or its combination with CVN on the SRS-based ASR framework (CMN+SRS and CMN+CVN+SRS) gave WERs which are comparable with the best ones achieved with the baseline system, using MAP adaptation, either on microphone array or lapel microphone signals. On the other hand, when there is overlapping speech from competing speakers (S_{12} and S_{123} scenarios), the combination of CMN and CVN on the SRS-based ASR framework (CMN+CVN+SRS) gave significant lower WERs, compared to the best ones achieved with the baseline system, using MAP adaptation, either on microphone array or lapel microphone signals. The SRS-based noise robust ASR framework has shown its relevance when combining with other standard noise robust techniques (CMN and CVN) to improve significantly microphone array speech recognition performance.

Table 1: Reference WERs (in %) computed by the ASR system, trained on original clean speech, tested on microphone array and lapel microphone signals, using MAP adaptation as noise robust technique. These results are extracted from [1].

	MAP adaptation	
Scenario	Microphone array	Lapel microphone
S_1	7.00	7.01
S_{12}	19.37	26.69
S_{123}	26.64	35.25

Table 2: WERs (in %) computed on microphone array signals using conventional ASR system, trained on original clean speech, and SRS-based ASR system, trained on SRS signals. CMN and CVN have been implemented on these two systems. The input speech signals are the same microphone array signals which are used to compute the WERs in Tab. 1.

	Noise robust technique					
Scenario	No technique	CMN	CMN+CVN	SRS	CMN+SRS	CMN+CVN+SRS
S_1	39.28	19.99	9.27	24.29	**7.51**	**7.16**
S_{12}	52.25	36.12	21.68	46.60	19.82	**17.16**
S_{123}	60.23	46.12	29.96	58.15	29.77	**23.29**

6 Conclusions

This paper investigates the possibility of improving microphone array speech recognition performance using cochlear implant-like SRS. Standard noise robust techniques, including CMN and CVN, have been implemented on a SRS-based noise robust ASR framework. The WERs obtained with the baseline system, using MAP adaptation, on microphone array and lapel microphone signals, were used as the references (see Tab. 1). Experiments, performed on MONC database [7], have shown that:

- When there is no overlapping speech, CMN+SRS and CMN+CVN+SRS gave WERs which are comparable with the lowest ones achieved with the baseline system, trained on original clean speech and tested, using MAP adaptation, on either microphone array or lapel microphone signals.
- When there is overlapping speech from competing speakers, CMN+CVN+SRS gave significantly lower WERs, compared to the lowest ones achieved with the baseline system, trained on original clean speech and tested, using MAP adaptation, on either microphone array or lapel microphone signals.
- CMN+SRS gave a lower WER compared to CMN+CVN on microphone array signals.

In fact, MAP adaptation is a robust technique but its implementation needs enough adaptation data to perform well. It has been shown that the performance of an ASR system using MAP adaptation depends heavily on the amount of adaptation data [12]. Therefore, it is inconvenient to implement MAP adaptation in real-time ASR systems. In this work, we have shown that implementing standard, less costly noise robust techniques (CMN and CVN) on the SRS-based ASR framework could help in achieving comparable or significantly lower WERs with microphone array signals, whenever there is no overlapping or overlapping speech, respectively, compared to those achieved with the baseline ASR system, using MAP adaptation. We hope to validate these results on a larger vocabulary ASR system such as one based on the AMI corpus [14] in our future work.

7 Acknowledgements

This work was supported by the Swiss National Science Foundation under the National Centre of Competence in Research (NCCR) on "Interactive Multimodal Information Management (IM2)". The authors gratefully thank the Swiss NSF for their financial support, and all project partners for a fruitful collaboration. More information about IM2 is available from the project web site www.im2.ch.

References

[1] D. C. Moore and I. A. McCowan, "Microphone array speech recognition: experiments on overlapping speech in meetings," in *Proc. IEEE ICASSP, April 06 - 10, Hong Kong, China*, Apr. 2003, vol. 5, pp. 497–500.

[2] M. Seltzer, "Bridging the gap: towards a unified framework for hands-free speech recognition using microphones arrays," in *Proc. HSCMA Hands-free speech communication and microphone arrays workshop*, May 2008, pp. 104–107.

[3] A. Stolcke, "Making the most from multiple microphones in meeting recognition," in *Proc. IEEE ICASSP 2011*, 2011, pp. 4992–4995.

[4] H. Gustafsson, S. E. Nordholm, and I. Claesson, "Spectral subtraction using reduced delay convolution and adaptive averaging," *IEEE Trans. on Speech and Audio Processing*, vol. 9, no. 8, pp. 799–807, Nov. 2001.

[5] R. V. Shannon, F.-G. Zeng, V. Kamath, J. Wygonski, and M. Ekelid, "Speech recognition with primarily temporal cues," *Science*, vol. 270, no. 5234, pp. 303–304, 1995.

[6] C.-T. Do, D. Pastor, and A. Goalic, "A novel framework for noise robust ASR using cochlear implant-like spectrally reduced speech," *Speech Communication*, vol. 54, no. 1, pp. 119–133, Jan. 2012.

[7] D. C. Moore and I. A. McCowan, "The multichannel overlapping numbers corpus (MONC)," 2003, http://www.cslu.ogi.edu/corpora/monc.pdf.

[8] G. Kubin and W. B. Kleijn, "On speech coding in a perceptual domain," in *Proc. IEEE ICASSP, March 15 - 19, Phoenix, AZ, USA*, Mar. 1999, vol. 1, pp. 205–208.

[9] T. S. Gunawan and E. Ambikairajah, "Speech enhancement using temporal masking and fractional Bark gammatone filters," in *Proc. 10th Australian Intl. Conf. on Speech Sci. & Tech., Dec. 8 - 10, Sydney, Australia*, Dec. 2004, pp. 420–425.

[10] C.-T. Do, D. Pastor, G. Le Lan, and A. Goalic, "Recognizing cochlear implant-like spectrally reduced speech with HMM-based ASR: experiments with MFCCs and PLP coefficients," in *Proc. Interspeech 2010, September 26-30, Makuhari, Chiba, Japan*, 2010, pp. 2634–2637.

[11] M. J. Taghizadeh, P. Garner, H. Bourlard, H. R. Abutalebi, and A. Asaei, "An integrated framework for multi-channel multi-source localization and voice activity detection," in *IEEE HSCMA Workshop*, 2011.

[12] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, X. Liu, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland, *The HTK book (for HTK version 3.4)*, Cambridge University Engineering Department, Cambridge, UK.

[13] P. N. Garner, "Cepstral normalisation and the signal to noise ratio spectrum in automatic speech recognition," *Speech Communication*, vol. 53, no. 8, pp. 991–1001, October 2011.

[14] S. Renals, T. Hain, and H. Bourlard, "Recognition and understanding of meetings: the AMI and AMIDA projects," in *Proc. IEEE ASRU 2007*, 2007, pp. 238–247.