ULTRASOUND-COUPLED SEMI-SUPERVISED NONNEGATIVE MATRIX FACTORISATION FOR SPEECH ENHANCEMENT

Tom Barker[⋆]*, Tuomas Virtanen*

Tampere University of Technology
Department of Signal Processing
Tampere, Finland
{Thomas.Barker, Tuomas.Virtanen}@tut.fi

Olivier Delhomme[†]

Université de Strasbourg
Télécom Physique Strasbourg
France
Olivier.Delhomme@etu.unistra.fr

ABSTRACT

We present an extension to an existing speech enhancement technique, whereby the incorporation of easily obtained Doppler-based ultrasound data, obtained from frequency shifts caused by a talker's mouth movements, is shown to improve speech enhancement results. Noisy speech mixtures were enhanced using semi-supervised nonnegative matrix factorisation (NMF). Ultrasound data recorded alongside the speech is transformed into the spectral domain and used additionally to audio in the mixture to be separated. Speech components are learned from a training set, whilst noise components are estimated from the mixture signal. We show that the ultrasound data can improve source-to-distortion ratios for the enhanced speech, relative to both the non-ultrasound NMF case and an established Wiener filter-based speech enhancement method.

Index Terms— Nonnegative Matrix Factorisation, Ultrasound, Acoustic Doppler Sensor, Source Separation

1. INTRODUCTION

Speech enhancement is an important research problem in automatic speech-recognition, hearing aids and telecommunication scenarios. The removal of noise from a transmitted speech signal can improve human communication, and automatic speech recogniser performance. Traditional techniques of spectral subtraction [1] and Wiener filtering [2] both require an estimate of the noise spectrum, to be most effective. This is generally obtained by speech activity detection being used to label portions of the signal as either speech or noise, or assuming that frames of the signal contains no speech activity, in order to derive a noise spectrum estimate.

Nonnegative matrix factorisation (NMF) is an established technique in sound source separation, and several variation and extensions exist which produce good performance in specific conditions. Separating a noisy speech signal into speech and noise allows noise removal and improved signal-to-noise ratios. NMF decomposes a signal into a sum of components which have time-varying activations and fixed spectra. Supervised NMF techniques generally learn the spectra of mixture components in advance, and estimate their temporal activations. Fully supervised NMF-based separation produces good separation performance where sufficient training data is available and where the noise model accurately fits the test data such as in [3]. Where a suitable noise model is not available, the performance of such approaches decreases. These scenarios are where semi-supervised approaches can offer improved performance. Generally, an effective semi-supervised approach uses fixed atoms for a single source, whilst noise atoms are obtained through iterative updates during factorisation, as in [4, 5].

This paper proposes the addition of ultrasound information to NMF-based audio source separation to aid in the removal of noise and enhance speech signal quality. Ultrasound has successfully been employed alongside audio data in improving performance of voice-activity detection [6], speaker recognition [7], and speech recognition [8]. Ultrasound is also ultilised in various other human-computer interfaces, as summarised in [9]. This paper details the first use of ultrasound data in an NMF audio-separation framework for speech enhancement.

Our approach employs hybrid atoms which incorporate both audio and ultrasound data for pre-training the speech components of a mixture, whilst estimating noise spectra. An ultrasound transmitter-receiver pair is used, which captures facial movements through Doppler shifts of the reflected ultrasound signal. It offers a low-hardware cost, intrinsically audio-noise-robust method of obtaining additional data correlated with the audio produced by a talker.

Doppler-shift based methods, as used in the proposed separation algorithm, rely on the movement of mouth and lips causing a shift in the reflected carrier frequency. Other multimodal approaches incorporate movement of objects to aid audio separation, such as through video as in [10], however this approach is more expensive both in terms of computation and hardware cost. The proposed method removes the requirement to extract features from the additional information stream, instead the captured data to be directly incorporated into the semi-supervised NMF framework.

2. ULTRASOUND DATA

Ultrasound data is captured simultaneously with audio. A single fixed frequency ultrasound carrier tone is transmitted at the talker's mouth and the reflected signal simultaneously received as in Figure 1. The ultrasound signal is produced by driving the transmitter of a tranceiver pair with a 40 kHz sinusoidal signal from a laboratory signal generator. The tranceiver pair has a narrowband response centred around 40kHz. Movements of the talker's lips and face cause a Doppler shift in the reflected signal, and modulations in frequency

⋆ The author would like to extend his thanks and appreciation to Nemanja Cvijanovic of Philips Research for his valuable advice and insight on the use of ultrasound transducers.

† Performed the work while visiting at Tampere University of Technology.

The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement number 290000 and Academy of Finland grant number 258708.

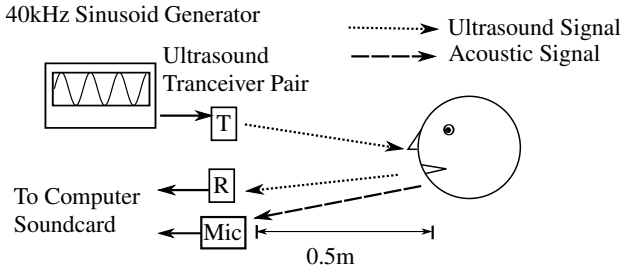

Fig. 1. Schematic diagram of the capture of acoustic and corresponding ultrasound Doppler data. T=Transmitter, R=Receiver, Mic=Audio Microphone.

around the carrier are produced, proportional to the normal velocity of the reflecting surface. In [7] this system is referred to as an 'acoustic Doppler sonar' (ADS). In this way, an ultrasound signal which is related to facial movements, hence, vocal sounds produced is captured.

2.1. Signal Capture

The transmitter and receiver are mounted on electronics prototyping 'veroboard' along with soldered input and output connections and 250 mV peak-to-peak sinusoid applied, which produces a sound pressure level of approximately 85 dB referenced to 20 μPa. The output of both the ultrasound receiver and simultaneously, a condenser microphone are captured by a computer soundcard at a samplerate of 96 kHz. Following capture, the ultrasound signal is demodulated to a centre frequency of 4.4 kHz, by multiplication with a digital sinusoid of frequency 35.6 kHz. The demodulated ultrasound signal is then lowpass filtered at 10 kHz and downsampled to 16 kHz. The audio data is also downsampled to 16 kHz to reduce computational complexity when performing the enhancement.

2.2. Data Representation

A sliding 1024 point Hamming window with 75% overlap is used to window both the time domain audio and (demodulated) ultrasound signals, with each frame undergoing a discrete Fourier transform (DFT) and the absolute value taken; effectively performing a short-time Fourier transform (STFT). Redundant complex conjugate data is discarded, leaving $L = 513$ bins of useful information. Due to the relatively narrowband nature of the demodulated ultrasound signal, not all bins of the 513-point DFT result are useful. In our implementation, $M = 32$ DFT bins of ultrasound data are retained from the DFT data, centred around the carrier frequency. The magnitude of the carrier frequency is significantly higher than the Doppler-shift-induced frequency variations, yet contains little useful information. The carrier frequency is therefore removed by setting corresponding frequency bins to zero to make pattern fitting to the Doppler regions (where information of interest exits) more effective (Figure 2).

To account for the longer-term low-frequency modulation trends of facial movement, several (temporally) successive STFT frames are concatenated to form each feature vector, for both the ultrasound and audio data. The number of frames taken in context is defined here as N. The operation is as in [11], where a sliding window of N frames is used to produce each frame of the modified output. The frames at the end of a signal are padded with zeros to reach correct dimensionality in the case that the column index exceeds the number of STFT frames. The sliding window concatenation operation is performed for the audio and ultrasound STFT data separately. The resulting matrices are combined, resulting in each observation vector having a length of $N(L + M)$. An example of the data representation is shown graphically in Figure 3. Where there are T observation frames, $\mathbf{X}$ is a matrix of dimensions $N(L + M) \times T$.

3. INTEGRATION OF ULTRASOUND INTO AN NMF FRAMEWORK

Nonnegative matrix factorisation represents a mixture signal as an additive sum of spectral atoms and their corresponding temporal activation. In our semi-supervised case, a dictionary matrix of speech atoms $\mathbf{B}_s$ is learned prior to factorisation, whilst noise dictionary matrix $\mathbf{B}_n$ is initialised randomly and updated during factorisation. The overall dictionary of atoms, $\mathbf{B}$ is comprised of $\mathbf{B}_s$ and $\mathbf{B}_n$ thus:

$$\mathbf{B} = [\mathbf{B}_s \mathbf{B}_n]. \tag{1}$$

The non-negative mixture signal, represented by magnitude spectrogram matrix $\mathbf{X}$ can then be modelled as matrix $\hat{\mathbf{X}}$, the product of dictionary, $\mathbf{B}$ with their corresponding weight matrix, $\mathbf{W}$ as:

$$\mathbf{X} \approx \hat{\mathbf{X}} = \mathbf{BW}. \tag{2}$$

3.1. Coupled Ultrasound Model

The audio STFT is modelled as the weighted sum of speech and noise atoms. The ultrasound STFT is modelled as the sum of coupled speech ultrasound atoms, having the same weight as the coupled audio portion, with no contribution from noise atoms.

In our case the matrix $\mathbf{X}$ consists of an audio and ultrasound portion, $\mathbf{X}_a$ and $\mathbf{X}_u$. The ultrasound portion is weighted by variable β, which defines its contribution to the mixture to be separated.

The dictionary matrix $\mathbf{B}$ can be divided into speech and noise, audio and ultrasound sections, with subscript s, n, a, u denoting speech, noise, audio and ultrasound respectively:

$$\mathbf{X} = \begin{bmatrix} \mathbf{X}_a \\ \beta\mathbf{X}_u \end{bmatrix} \approx \hat{\mathbf{X}} = \begin{bmatrix} \mathbf{B}_{sa} & \mathbf{B}_{na} \\ \beta\mathbf{B}_{su} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{W}_s \\ \mathbf{W}_n \end{bmatrix} \tag{3}$$

where $\mathbf{W}_s$ and $\mathbf{W}_n$ are the speech and noise atom weights respectively. There is no ultrasound data used in the noise model, instead, this part of the matrix is filled with zeros.

Here, we model a mixture where there is no ultrasound present in the background noise. We make the assumption that in application of such a method, noise-based interference at the tranceiver response frequency is minimal due to the narrowband nature of the ultrasound signal. Indeed, in [6] the effects of spurious background noise on such a Doppler based system were evaluated and had an inconsequential effect.

3.2. NMF Algorithm

The NMF algorithm aims to minimise the Kullback-Leibler (KL) divergence between $\hat{\mathbf{X}}$ and $\mathbf{X}$, whilst constraining sparsity of the weights as in [12]. Sparsity constraints have been shown in some cases to be able to produce better separation within an NMF framework by producing more 'meaningful' data representations, penalising non-zero terms in $\mathbf{W}$. That is, the spectra which typically constitute a sound source are only active for a small portion of the temporal frames present. The update equations are described here are applied on $\mathbf{B}$ and $\mathbf{W}$, to factorise $\hat{\mathbf{X}}$. Updates are performed on $\mathbf{B}_n$, $\mathbf{W}_s$ and $\mathbf{W}_n$, updating both the weights and noise atoms during each iteration. Speech atoms remain fixed.

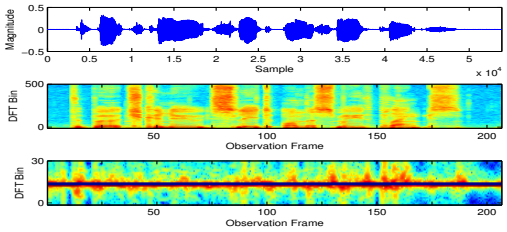

Fig. 2. Original audio signal (top), audio signal spectrogram (middle) and corresponding processed ultrasound spectrogram centred on the data of interest with carrier frequency removed (bottom).

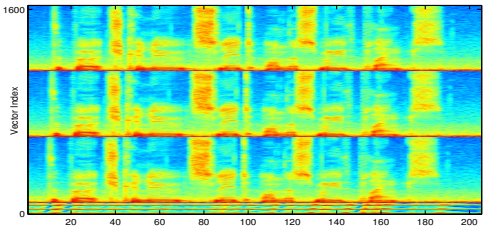

Fig. 3. Example matrix $\mathbf{X}$ reshaped with $N = 3$. The upper portion is the reshaped audio contribution, whilst the narrow band at the bottom of the figure is the reshaped ultrasound contribution.

The weights are updated using Equation (4) after matrix $\mathbf{W}$ has been initialised to value 1 over its entirety.

$$\mathbf{W} \leftarrow \mathbf{W} \otimes \frac{\nabla c^{-}(\mathbf{B}, \mathbf{W})}{\nabla c^{+}(\mathbf{B}, \mathbf{W})} \tag{4}$$

where:

$$[\nabla c^{-}(\mathbf{B}, \mathbf{W})]_{j,t} = [\mathbf{B}' \frac{\mathbf{X}}{\mathbf{BW}}]_{j,t} + \alpha \frac{w_{j,t}\sqrt{T}\sum_{i=1}^{T} w_{j,i}}{(\sum_{i=1}^{T} w_{j,i}^2)^{3/2}} \tag{5}$$

$$[\nabla c^{+}(\mathbf{B}, \mathbf{W})]_{j,t} = [\mathbf{B}'\mathbf{1}]_{j,t} + \alpha \frac{1}{\sqrt{\frac{1}{T}\sum_{i=1}^{T} w_{j,i}^2}} \tag{6}$$

and where α is a sparsity weight, $\mathbf{1}$ is a matrix of all ones, with the same dimensionality as $\mathbf{X}$ and T is the number of observation frames. Matrix row indices are denoted as j and column indices t or i. The first terms of Equations (5) and (6) correspond to KL-divergence whilst the last terms correspond to sparsity.

The noise portion $\mathbf{B}_{na}$ of the spectral atom matrix $\mathbf{B}$ is updated using:

$$\mathbf{B}_{na} \leftarrow \mathbf{B}_{na} \otimes \frac{\frac{\mathbf{X}_a}{\mathbf{BW}}\mathbf{W}_n{}'}{\mathbf{1}\mathbf{W}_n{}'}. \tag{7}$$

Both the full weights matrix $\mathbf{W}$ and the audio portion noise spectra atoms $\mathbf{B}_{na}$ are updated iteratively, alternating between Equation (4) applied to $(\mathbf{B}, \mathbf{W})$ and Equation (7) applied to $\mathbf{B}_{na}$.

3.3. Prevention of Overfitting

In semi-supervised NMF approaches where noise atoms are updated, there is the possibility that noise atoms can model the speech, or entire mixture, rather than the noise contribution. When applying update equations until convergence is reached, overfitting of the noise basis atoms can occur. Indeed in [2] it is noted that separation performance decreased with increasing number of iterations. A compromise must be reached between updating the weights matrix and overfitting the noise atoms. In our trials we evaluated the effects of increasing numbers of iterations of update equations applied to both weights and noise atoms.

Stacking successive frames in the factorisation model also reduces overfitting, since shorter unstacked noise atoms would more rapidly converge to represent speech. Higher numbers of noise atoms also make overfitting more likely.

3.4. Signal Reconstruction

Following factorisation in the spectral domain, the separated audio components must be reconstructed in the time domain. For this, a Wiener filter is produced from the audio portion of the factorised matrices. Ultrasound information is discarded, and a stacked spectrogram for speech factors is generated by:

$$\tilde{\mathbf{Y}}_s = \mathbf{B}_{as}\mathbf{W}_s \tag{8}$$

and similarly for noise

$$\tilde{\mathbf{Y}}_n = \mathbf{B}_{an}\mathbf{W}_n. \tag{9}$$

These matrices are reshaped to dimensions of $L \times T$, by applying an inverse to the concatenation operation described in Section 2.2. Columns are unstacked, and contributions placed additively into their original position in the audio STFT. Estimates of $\mathbf{Y}_s$ and $\mathbf{Y}_n$ are obtained from averaging overlapping windows as in [11]. The speech contribution $\mathbf{S}$, to the original audio mixture spectrogram $\tilde{\mathbf{X}}_a$ can be obtained by applying the Wiener-like filter:

$$\mathbf{S} = \tilde{\mathbf{X}}_a \otimes \frac{\mathbf{Y}_s}{\mathbf{Y}_s + \mathbf{Y}_n} \tag{10}$$

The time-domain speech signals is then obtained from $\mathbf{S}$ by combination with the original phase from $\tilde{\mathbf{X}}_a$, applying an inverse STFT, and overlap-add reconstructing the resulting frames.

4. EXPERIMENTAL EVALUATION

The performance of the proposed method was evaluated using the source-to-distortion ratio (SDR) metric, as defined in the BSS Toolkit for evaluating source separation [13]. The enhancement approach was compared to semi-supervised NMF without ultrasound as well as a two-step Wiener-filter spectral subtraction method [2]. Noise was removed with the proposed method, and the SDR of the enhanced speech was measured.

4.1. Producing the Test Mixtures

Test mixtures were created by mixing a single captured ultrasound utterance with noise data. Non-stationary noise data consisted of random portions of the background noise portion of the CHiME corpus [14] (realistic environmental noise).

Speech data was produced by a native British English speaker reading aloud Harvard sentences [15]. All sentences were recorded in one sitting, with the ultrasound transducer and microphone placed 50 cm away from the talker's face, and the talker remaining stationary.

100 Harvard sentences were mixed with portions of the noise data, trimmed to the length of the speech utterance. All mixtures were normalised to have the same root-mean-squared (RMS) value for both speech and noise. Mixtures were only created for the audio portion of the signal, whilst the ultrasound signal was copied from the speech signal. The CHiME dataset was also used in evaluation of

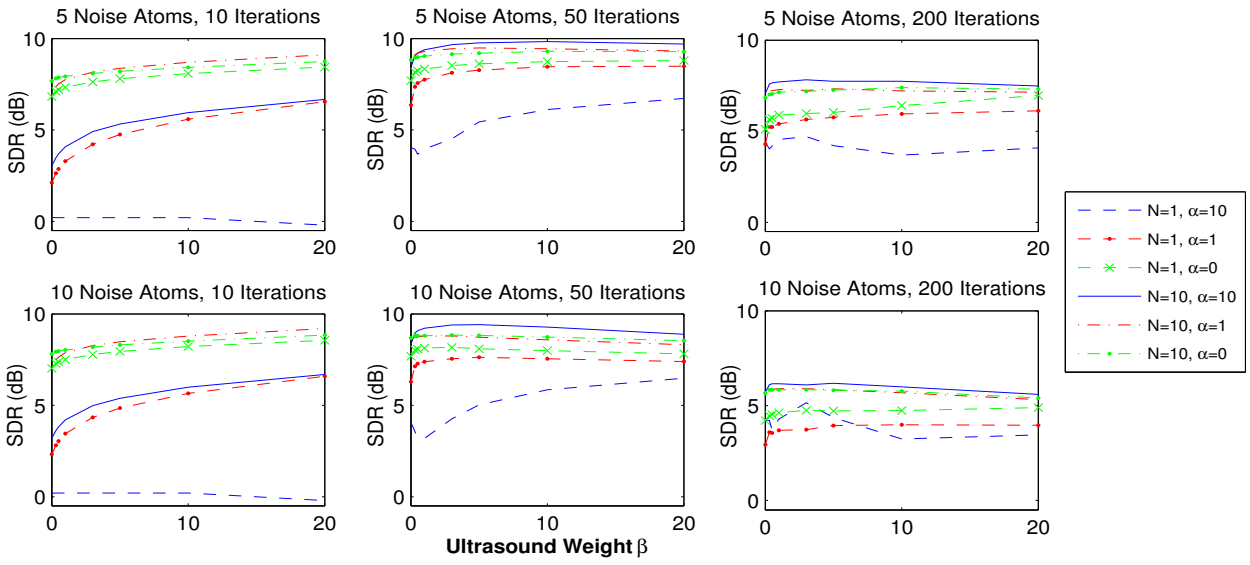

Fig. 4. Average speech enhancement SDRs for 10, 50 and 200 iterations of updates with 5 and 10 noise atoms and varying sparsity weight α, relative to ultrasound weight, β. Each panel shows the effect of varying sparsity weight and number of stacked frames, N, for a fixed number of iterations and noise atoms.

semi-supervised NMF of speech/noise mixtures in [5], although our test mixtures had a normalised SNR of 0 dB whereas the mixtures are unscaled in [5], as well as using a different dictionary atom length, making direct comparison of SDR figures obtained difficult.

4.2. Obtaining the Dictionary Atoms

Cross validation over 100 uterances was employed with a single utterance forming the test-mixture trial subset, whilst dictionary atoms were obtained from the remaining 99 utterances (the test subset). The atoms were produced by subjecting each training utterance to the same processing that was used to produce the test mixture, that is, STFT and stacking, with the same parameters as to produce test mixture matrix $\mathbf{X}$. This resulted in a dictionary of roughly 20,000 pre-learned atoms for each trial. The audio portion of the noise atoms was randomly initialised with absolute value of Gaussian noise, which was updated via the equations in Section 3.2, whilst the ultrasound portion (not updated) was initialised with a very small positive random values (of the order 10^{-12}), to prevent divide-by-zero errors.

4.3. Test Conditions

The proposed method was evaluated across the 100 test mixtures for different values of frame stacking, N, and factorisation sparsity cost α. The mean SDR value over 100 trials was obtained for each ultrasound weight β. Values of 0, 1 and 10 were used for α and 1, 5, and 10 for N. Noise dictionaries of size $M = 5$ and $M = 10$ atoms were trialled, as were performing 10, 50 or 200 iterations of update equations. The ultrasound portion of $\mathbf{X}$ was weighted with a parameter β, specifically β =0, 0.3, 0.5, 1, 3, 5, 10, 20. The case of $\beta = 0$ was verified to produce numerically identical to exclusion of the ultrasound data altogether.

4.4. Experimental Results

The results of interest are presented in Figure 4. In all cases tested, increasing number of stacked frames, N, produced better enhancement. The case of $N = 5$ is not presented, as results produced were inferior to the $N = 10$ case. As the number of iterations increased, overall performance varied. It is shown that increasing from 10 to 50 interations increases performance, but when 200 iterations are applied (convergence is reached), performance generally decreases compared to the 50 iteration case. The results show that for the mixtures under test, the incorporation of ultrasound data improves NMF-based enhancement performance. Even in the best performing cases, over 1dB additional enhancement performance is achieved as a result of inclusion of ultrasound data.

Greatest performance is achieved for high sparsity weight for a high N, and lower number of noise atoms. The effects of sparsity also vary with number of update iterations, and at low number of iterations the performance difference between $M = 5$ and $M = 10$ noise atoms is negligible. The Wiener filtering achieved an average enhancement of 3.54 dB, so is surpassed by the NMF-based approach in almost all practical parameter combinations.

5. CONCLUSIONS AND DISCUSSIONS

This paper proposed a way to incorporate information about facial movements measured with ultrasound into a semi-supervised NMF based speech enhancement framework. The Doppler shifts caused by reflected ultrasound are used to produce spectral features which can be employed alongside audio spectrograms. The results showed increased enhancement with the addition of ultrasound data, and provided insight into how algorithm parameters affect enhancement.

The effects of overfitting the noise atoms should studied further, as the intereaction between number of updates, sparsity constaints and stacking of temporal frames is complicated. It is likely that where the performance decreased despite more convergence of the cost function, that noise atoms had increasingly approximated the speech contributions. It is possible that improved results obtained through stacking is a combination of both decreased overfitting and inclusion of greater temporal context.

6. REFERENCES

[1] S. Boll, "Suppression of Acoustic Noise in Speech Using Spectral Subtraction," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 27, no. 2, pp. 113–120, 1979.

[2] C. Plapous, C. Marro, and P. Scalart, "Improved Signal-to-Noise Ratio Estimation for Speech Enhancement," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 6, pp. 2098–2108, 2006.

[3] T. Virtanen, J.F. Gemmeke, and B. Raj, "Active-Set Newton Algorithm for Overcomplete Non-Negative Representations of Audio," *IEEE Transactions on Audio, Speech and Language Processing*, 2013.

[4] G J. Mysore and P. Smaragdis, "A Non-Negative Approach to Semi-supervised Separation of Speech from Noise with the use of Temporal Dynamics," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2011, pp. 17–20.

[5] C. Joder, F. Weninger, F. Eyben, D. Virette, and B. Schuller, "Real-Time Speech Separation by Semi-supervised Nonnegative Matrix Factorization," in *Latent Variable Analysis and Signal Separation*, vol. 7191 of *Lecture Notes in Computer Science*, pp. 322–329. Springer Berlin Heidelberg, 2012.

[6] K. Kalgaonkar, R. Hu, and B. Raj, "Ultrasonic Doppler Sensor for Voice Activity Detection," *IEEE Signal Processing Letters*, vol. 14, no. 10, pp. 754–757, 2007.

[7] K. Kalgaonkar and B. Raj, "Ultrasonic Doppler Sensor for Speaker Recognition," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2008, pp. 4865–4868.

[8] B. Zhu, "Multimodal Speech Recognition with Ultrasonic Sensors," M.S. thesis, Massachusetts Institute of Technology, 2007.

[9] B. Raj, K. Kalgaonkar, C. Harrison, and P. Dietz, "Ultrasonic Doppler Sensing in HCI," *IEEE Pervasive Computing*, vol. 11, no. 2, pp. 24–29, 2012.

[10] A.L. Casanovas, G. Monaci, P. Vandergheynst, and R. Gribonval, "Blind Audiovisual Separation Based on Redundant Representations," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2008, pp. 1841–1844.

[11] J.F. Gemmeke, T. Virtanen, and A. Hurmalainen, "Exemplar-Based Sparse Representations for Noise Robust Automatic Speech Recognition," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 19, no. 7, pp. 2067–2080, 2011.

[12] T. Virtanen, "Monaural Sound Source Separation by Nonnegative Matrix Factorization With Temporal Continuity and Sparseness Criteria," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 3, pp. 1066 –1074, march 2007.

[13] E. Vincent, R. Gribonval, and C. Fevotte, "Performance Measurement in Blind Audio Source Separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 4, pp. 1462 –1469, July 2006.

[14] H. Christensen, J. Barker, N. Ma, and P D. Green, "The CHiME Corpus: a Resource and a Challenge for Computational Hearing in Multisource Environments," in *Proceedings of INTERSPEECH2010*. 2010, pp. 1918–1921, ISCA.

[15] E.H. Rothauser et al., "IEEE Recommended Pratice for Speech Quality Measurements," *IEEE Transactions on Audio and Electroacoustics*, vol. 17, no. 3, pp. 225–246, 1969.