A TWO-CHANNEL REVERBERATION SUPPRESSION SCHEME BASED ON BLIND SIGNAL SEPARATION AND WIENER FILTERING

Andreas Schwarz, Klaus Reindl, Walter Kellermann

Multimedia Communications and Signal Processing
University of Erlangen-Nuremberg
Erlangen, Germany
{schwarz, reindl, wk}@lnt.de

ABSTRACT

In this paper, we apply a blind signal extraction scheme for two microphones to the problem of dereverberation. The system consists of a blocking matrix that cancels the target signal as well as reverberated components up to a certain time lag, thus obtaining a reference not only for noise and interference, but also for late reverberation, which can then be suppressed with a Wiener filter, while leaving early reverberation components largely intact. The performance is assessed in terms of recognition rate of an automatic speech recognizer trained on clean speech, using sentences from the GRID corpus convolved with measured room impulse responses. We show that the system, although primarily developed for noise and interference suppression in low SNR conditions, can significantly suppress reverberation and thereby improve recognition results.

***Index Terms*—** Dereverberation, blind source separation, source extraction, speech enhancement

1. INTRODUCTION

Many emergent applications require the distant recording of speech signals, e.g., video telephony, where the microphones are usually placed near the camera, or voice control in home automation, where the user should be able to be untethered while controlling applicances.

Distant-talking speech interfaces introduce a number of additional challenges. One of the main factors for deteriorating speech intelligibility and speech recognition performance in such scenarios is reverberation. While the early part of a room impulse response only causes a coloring of the signals, late reverberation causes a temporal smearing of speech features that affects both speech intelligibility and the performance of automatic speech recognition (ASR) systems.

Many algorithms have been proposed for speech dereverberation. They can be differentiated in two categories: inverse filtering algorithms (e.g., MINT [1], linear prediction-based deconvolution [2] or TRINICON [3]), and algorithms that estimate and suppress reverberation, e.g., with spectral subtraction or Wiener filtering (see, e.g., [4]). For an overview of the state of the art, see [5].

In this paper, we describe a two-channel blind signal extraction (BSE) algorithm that was originally developed for the suppression of point-like interferers and diffuse noise, and investigate its performance for dereverberation in terms of speech recognition accuracy. The algorithm consists of a blind source separation (BSS) system for estimating undesired signal components (here: reverberation) and a Wiener filter for suppressing these components in the microphone signals. The BSS algorithm is based on the TRINICON framework [6]. Our BSE system was already presented and investigated for noise and interference reduction in various applications [7, 8, 9]. Other researchers have published similar approaches based on different BSS algorithms, e.g., in [10], but to our best knowledge these concepts have not been applied to reverberation suppression. Although a combination of BSS-based interference and reverberation suppression was presented in [11], there, a separate late reverberation estimator based on a model of the late impulse response was used. In our algorithm, no estimation or model of the impulse response is needed in order to reduce reverberation; instead, the algorithm can estimate and suppress late reverberation components in the same way as other undesired signal components.

The paper is organized as follows: our blind signal extraction scheme is briefly reviewed in Sect. 2, the experimental setup and results are presented in Sect. 3, and Sect. 4 concludes the paper.

2. BSS-BASED SOURCE EXTRACTION

The general problem addressed by our BSE system is illustrated in Fig. 1. A mixture of a desired signal s_1 and $Q-1$ interfering signals $s_2 \dots s_Q$ is recorded with $P = 2$ omnidirectional microphones. The signal path between source q, $q = [1, \dots, Q]$ and microphone p, $p \in \{1, 2\}$ is modeled as a convolution of the signals with room impulse responses h_{qp}, i.e., we record a convolutive mixture. In general, additive noise of unknown coherence $n_{\mathrm{b},p}, p \in \{1, 2\}$ may also be present. Signals are represented in the discrete time domain, sampled with a sampling rate f_s.

For the investigation of the dereverberation properties of the algorithm in this paper, we assume that no additive noise and only the desired source s_1 are present, with the desired source located at approximately 0° (broadside direction). In practice, this constraint on the target source position can be removed by combining the source extraction scheme with a localization algorithm [7].

We furthermore define $h_{qp,\mathrm{E}}$ and $h_{qp,\mathrm{L}}$ as the early and late parts of the impulse responses, respectively, with the time T_E defining the instant when the late reverberation starts:

$$h_{qp,\mathrm{E}} = h_{qp}(0 \dots T_\mathrm{E} f_\mathrm{s} - 1) \quad (1)$$

$$h_{qp,\mathrm{L}} = h_{qp} - h_{qp,\mathrm{E}} \quad (2)$$

The structure of the BSE unit is shown in Fig. 2. The main components are a BSS-based blocking matrix that estimates a noise reference from the microphone signals by canceling the desired source, and a noise suppression filter driven by this noise reference.

In the following sections, these two components of the system are described in more detail, and we will show how this scheme,

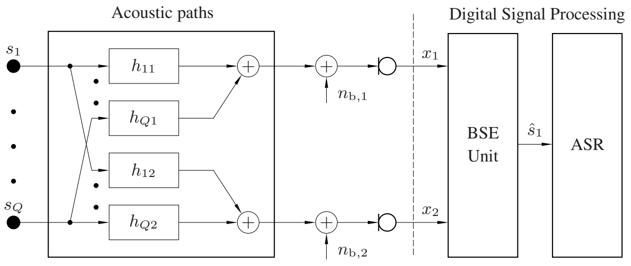

Fig. 1. Signal model for source extraction with speech recognition

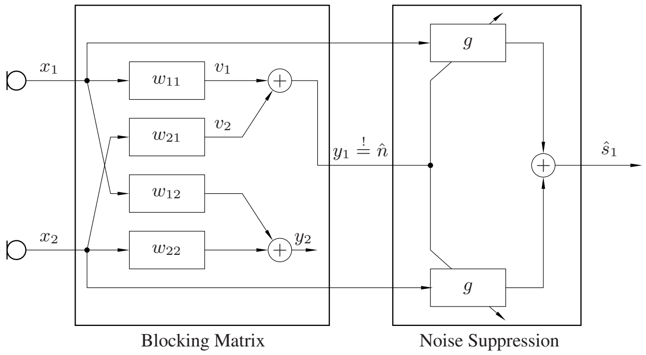

Fig. 2. Realization of the BSE unit

originally designed for noise and interference suppression, can be generalized for reverberation suppression.

2.1. Estimation of Undesired Signal Components

The aim of the blocking matrix is to suppress the desired signal s_1 contained in the recorded microphone signals as strongly as possible, in order to continuously provide a good reference for all the undesired signal components acquired by the microphones. The corresponding noise components in x_1 and x_2 can then be suppressed with a time-varying filter g.

We use a BSS system based on the TRINICON framework [6]. The algorithm adapts the demixing filters w_{pq}, $q \in \{1, 2\}$ (Fig. 2) with the aim of minimizing mutual information between outputs, and thereby separating independent components of the recorded mixture. However, with two microphones, this approach can only separate two sources (determined case), which is why we do not use the BSS algorithm directly for the extraction of the desired source, but as a noise estimator. To separate all noise components from the target signal in the general underdetermined case, the BSS algorithm is extended by a directional constraint that forces a spatial null towards the direction of the desired source, while suppressing correlated components from other directions (i.e., reflections) as well [7]. Thus, we use only one of the output signal paths of the BSS demixing system, the interference canceler formed by w_{11} and w_{21} (Fig. 2).

The directional constraint is implemented by extending the BSS cost function J_{BSS} by the geometrical constraint J_{C} [12, 7]:

$$J_{\mathrm{DirBSS}} = J_{\mathrm{BSS}} + \eta_{\mathrm{C}} J_{\mathrm{C}}, \tag{3}$$

$$J_{\mathrm{C}} = \parallel w_{11}(k) + w_{21}(k - \tau_\phi) \parallel^2, \tag{4}$$

with the TDOA τ_ϕ of the desired signal (here: $\tau_\phi = 0$ as it is assumed that the desired source is located in broadside direction). It can be seen that, when the demixing filters w_{11} and a τ_ϕ-shifted version of w_{21} cancel each other, this constraint becomes zero. The weight η_{C} controls the relative importance of the directional constraint [7].

By setting the directional constraint so that the desired source is suppressed in output 1 of the demixing system, we obtain the noise estimate as:

$$\hat{n} = y_1 = v_1 + v_2 = w_{11} * x_1 + w_{21} * x_2. \tag{5}$$

So far, the generic functionality of the blocking matrix has been described. In what follows, it is shown how this directional BSS-based concept can be used to estimate late reverberation, i.e., to separate early reflections from late reverberation. The reason why we specifically target late reverberation is, that it has been shown that the suppression of early reverberation (up to approx. 50 ms) is detrimental to speech recognition performance [13].

As shown in [7], directional BSS as a blocking matrix can cancel the direct path and correlated components, i.e., reverberation. As a consequence, reverberation components of the target signal are not contained in the noise reference, and therefore not suppressed, so that distortion of the target signal is avoided. By truncating the BSS demixing filters to a length corresponding to the early reverberation part up to the threshold $T_{\mathrm{E}} \approx 50\,\mathrm{ms}$, the BSS algorithm can only cancel the early reverberation up to T_{E}, and thus we obtain a reference $\hat{n}$ for the late reverberation.

An ideal solution for the demixing filters, i.e., filters that ideally equalize and thereby cancel the desired signal and its early reverberation components, is given by:

$$w_{11} = h_{\mathrm{E},12} \tag{6}$$

$$w_{21} = -h_{\mathrm{E},11}. \tag{7}$$

2.2. Suppression of Undesired Signal Components

For suppression of the undesired signal components (here: late reverberation), the signals are divided into subbands using an oversampled DFT filterbank. In the following, subband signals are denoted with the corresponding uppercase letters.

For each processing block k and subband γ, a frequency-domain filter weight $G^{(\gamma)}[k]$ is computed based on Wiener filtering with an overestimation factor μ and a spectral floor $G_{\min}$ [14]:

$$G^{(\gamma)}[k] = \max\left(1 - \mu \frac{1}{\hat{\mathrm{SNR}}^{(\gamma)}_{\mathrm{est}}}, G_{\min}\right). \tag{8}$$

The frequency-dependent SNR of the microphone signals is estimated based on the reference $\hat{n}$ and the filtered microphone signals v_1 and v_2:

$$\hat{\mathrm{SNR}}^{(\gamma)}_{\mathrm{est}} = \frac{\frac{1}{2}(S^{(\gamma)}_{v_1}[k] + S^{(\gamma)}_{v_2}[k])}{S^{(\gamma)\prime}_{\hat{n}}[k]}, \tag{9}$$

where $S^{(\gamma)\prime}_{\hat{n}}[k]$ is the corrected (see below) PSD (power spectral density) of the noise estimate at the BSS output, and $S^{(\gamma)}_{v_p}[k], p \in \{1, 2\}$ are the PSDs of the BSS-filtered microphone signals. The power spectral densities are estimated recursively with a forgetting factor λ, with $0 < \lambda < 1$.

An additional correction of the noise PSD is performed with the coherence of the noise at the microphones:

$${S_n}^{(\gamma)\prime}[k] = \frac{{S_n}^{(\gamma)}[k]}{2(1 + \Re\{\Gamma^{(\gamma)}\})} \tag{10}$$

This correction is necessary in order to obtain a noise PSD estimate that can be applied to the individual microphone channels from the combined noise PSD, and is discussed in detail in [9].

If we approximate the noise field as spherically isotropic [15] and the microphones as omnidirectional, and assume that the BSS demixing filters do not change the coherence of the noise component, we obtain the theoretical coherence function [16]

$$\Gamma^{(\gamma)}_{\text{diffuse}} = -\,\text{sinc}\left(\frac{2\pi f^{(\gamma)} d}{c}\right),\ \text{sinc}(\cdot) = \frac{\sin(\cdot)}{\cdot}, \qquad (11)$$

where $f^{(\gamma)}$ is the center frequency of the subband γ, d is the microphone spacing and c is the speed of sound.

As shown in Fig. 2, the noise reduction filter is applied to both microphone signals, which are then added to yield the output signal $\hat{s}_1$ (after resynthesizing the time-domain signal) as an estimate of the desired source signal s_1 with significantly reduced late reverberation:

$$\hat{S}^{(\gamma)}_1[k] = G^{(\gamma)}[k] \cdot (X^{(\gamma)}_1[k] + X^{(\gamma)}_2[k]). \qquad (12)$$

3. EXPERIMENTS

We evaluate the performance of our BSE algorithm for dereverberation based on speech recognition accuracy with a speech recognizer trained on clean speech. Since noise and interference suppression performance was already investigated in previous papers [8, 9], we focus only on the dereverberation aspect here.

3.1. Signals and Setup

We use clean speech sentences from the GRID corpus [17] as source signals. The GRID corpus consists of 34000 sentences with a simple syntax, spoken by 34 different speakers. For the evaluation, we use a test set of 2000 randomly selected sentences from the corpus.

To create the reverberated signals, we use impulse responses from a lecture hall with a reverberation time $T_{60} \approx 900\,\text{ms}$ and a critical distance $d_c \approx 0.9\,\text{m}$. The impulse responses were measured using maximum length sequences and truncated to 10000 samples at the sampling rate of $f_s = 16\,\text{kHz}$. Two scenarios are considered, one where the speaker-microphone array distance is $l = 2\,\text{m}$, and one where $l = 1\,\text{m}$. The two microphones are omnidirectional and placed 8.4 cm apart.

Fig. 3 compares the theoretical coherence of spherically isotropic noise and the actual coherence computed from the estimated late reverberation component. We can see that the assumption made in Sect. 2.2 matches the data quite closely.

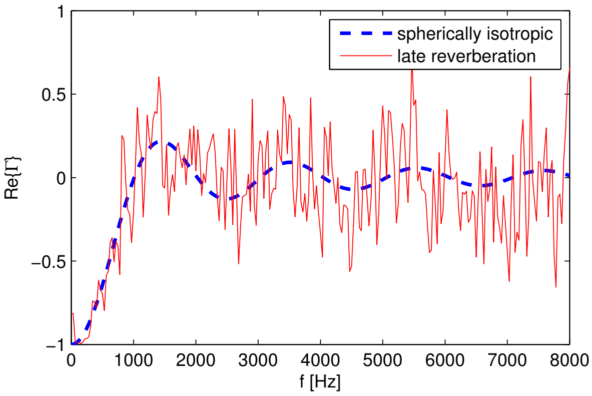

Fig. 3. Coherence for spherically isotropic noise, and actual coherence of late reverberation estimate ($l = 2\,\text{m}$)

3.2. Dereverberation

For the BSS-based blocking matrix, we compute a set of fixed BSS demixing coefficients that are obtained by concatenating multiple reverberated GRID sentences and performing the BSS adaptation. The demixing filter length is set to 800 samples, according to the defined late reverberation threshold $T_\text{E} = 50\,\text{ms}$. Additionally, to determine the maximum performance achievable by this system structure, we also investigate using ideal demixing filters with length 800 (according to Eq. 6) instead of BSS-adapted filters for the blocking matrix.

The frequency-domain Wiener filter is implemented with a polyphase filterbank with a filter length of 1024, 512 complex-valued subbands, and a downsampling rate of 128. The overestimation factor μ is set to 1.5, the maximum suppression gain G_{min} is set to 0.12. These values were empirically found to yield close to maximum recognition rates both with the directional BSS and the ideal demixing filters.

For further comparison, we investigate the performance of a Wiener filter based on the true late reverberation obtained by convolving the clean signals with only the late part (according to T_E) of the measured impulse responses. We use a common Wiener filter for both channels with the same parameters as in the BSE system, however with an overestimation factor $\mu = 1.2$ and without the coherence correction.

As a measure for the amount of late reverberation in the processed and unprocessed signals, we define a signal-based early to late reverberation ratio:

$$ELRR_{\mathcal{P}} = \frac{||\mathcal{P}(s_1 * h_\text{E})||^2}{||\mathcal{P}(s_1 * h_\text{L})||^2}, \qquad (13)$$

where $\mathcal{P}$ is the linear operator that describes the effect of the dereverberation operation on the early and late reverberation components of the signal s_1.

As an isolated measure for the success of the blocking matrix, we define the target suppression (here: the amount of early reverberation suppression) obtained by the blocking matrix:

$$TS_{\text{BM}} = \frac{ELRR_{x_1+x_2}}{ELRR_{\hat{n}}} = \frac{\frac{||s_1 * h_{11,\text{E}} + s_1 * h_{12,\text{E}}||^2}{||s_1 * h_{11,\text{L}} + s_1 * h_{12,\text{L}}||^2}}{\frac{||s_1 * h_{11,\text{E}} * w_{11} + s_1 * h_{12,\text{E}} * w_{21}||^2}{||s_1 * h_{11,\text{L}} * w_{11} + s_1 * h_{12,\text{L}} * w_{21}||^2}}. \qquad (14)$$

To quantify the amount of distortion of the desired signal components caused by the Wiener filter, we define the signal distortion

$$SD_{\mathcal{P}} = \frac{||\mathcal{P}(s_1 * h_\text{E}) - s_1 * h_\text{E}||^2}{||s_1 * h_\text{E}||^2}, \qquad (15)$$

where, again, $\mathcal{P}$ represents the dereverberation operation.

3.3. Automatic Speech Recognizer

In order to assess the improvement obtained by the proposed late reverberation suppression scheme, we perform speech recognition experiments with a recognizer based on Pocketsphinx [18]. A finite-state language model is defined according to the structure of the GRID sentences. The recognizer uses triphone HMMs with 3 states per model, 8 Gaussian output densities per state, and a total number of 600 tied states, with features based on 13 mel-frequency cepstral coefficients (MFCCs) with velocity and acceleration. The acoustic model is trained using 32000 clean speech sentences from the GRID corpus (the 2000 sentences from the test set are not used for training). No speaker-specific adaptation is performed.

For the recognition output, an accuracy score is computed in the same way as in the CHiME challenge [19], where only two words within the sentence (a letter and a digit) are considered.

		TS_{BM}	*ELRR*	accuracy	*SD*
	clean	-	∞	93.3%	-
l=1 m	sum	-	8.99 dB	81.0%	-
	BSE, proposed	**13.77 dB**	**12.47 dB**	**87.7%**	**-14.54 dB**
	BSE w. ideal demix.	∞	11.76 dB	88.0%	-16.86 dB
	WF w. ideal ref.	-	13.75 dB	90.5%	-18.75 dB
l=2 m	sum	-	3.93 dB	56.4%	-
	BSE, proposed	**5.79 dB**	**7.53 dB**	**73.9%**	**-9.32 dB**
	BSE w. ideal demix.	∞	7.37 dB	77.4%	-11.57 dB
	WF w. ideal ref.	-	10.08 dB	86.5%	-13.07 dB

Table 1. Experimental results: BSE target suppression, early to late reverberation ratio, recognition accuracy and signal distortion for clean, reverberated and processed signals.

3.4. Results

Table 1 summarizes the results of the experiments in both scenarios. The baseline recognition rate is obtained by applying the recognizer to the sum of both microphone signals. It can be seen that the proposed BSS-based reverberation suppression scheme improves the recognition accuracy significantly relative to the baseline and comes close to the performance obtained by using ideal demixing filters for the noise estimate. Note that the $ELRR$ improvement is similar in the cases of BSS and ideal demixing filters; however, using the latter causes a significantly lower signal distortion, due to the perfect target suppression of the blocking matrix.

Using an ideal reference of late reverberation components would further improve the result; this is because, although the ideal demixing filter perfectly cancels the early reverberation, it only provides a spatially filtered noise reference that has to be normalized again using the coherence assumption, and furthermore, because the demixing filters cause a temporal distortion of the noise reference and the microphone signals which are the basis for the suppression filter computation.

Note that the signals were free of noise and interference, the improvement in recognition rate is therefore solely based on the reduction of late reverberation. However, since the system is based not on estimating or modeling the late reverberation components directly, but suppressing the early reverberation components to obtain a reference of all undesired signal components, the presence of noise and interference does not impact dereverberation performance of the algorithm.

4. CONCLUSION

A two-channel source extraction scheme based on directional BSS for noise estimation and Wiener filtering for noise suppression was presented and applied to the reduction of late reverberation. We have shown that by using a BSS algorithm to cancel direct path and early reverberation components of a target signal, we can obtain a reference for not only noise and interference, but also late reverberation, which can then be suppressed with a Wiener filter, allowing a significant reduction of late reverberation and a corresponding improvement in speech recognition accuracy of an ASR system. While algorithms targeted only at reverberation may achieve still better results, this confirms the versatility of the blind source extraction approach for the reduction of arbitrary undesired signal components. We would like to emphasize that this approach is in general valid for all BSS algorithms that can achieve cancellation of direct path and early reverberation components of a target signal.

5. REFERENCES

[1] M. Miyoshi and Y. Kaneda, "Inverse filtering of room acoustics," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 36, no. 2, pp. 145–152, Feb. 1988.

[2] M. Delcroix, T. Hikichi, and M. Miyoshi, "Precise dereverberation using multichannel linear prediction," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 15, no. 2, pp. 430–440, Feb. 2007.

[3] H. Buchner, R. Aichner, and W. Kellermann, "Trinicon: a versatile framework for multichannel blind signal processing," in *Proc. ICASSP*, May 2004.

[4] E. A. P. Habets, *Single- and Multi-Microphone Speech Dereverberation using Spectral Enhancement*, Ph.D. thesis, Technische Universiteit Eindhoven, 2007.

[5] P.A. Naylor and N.D. Gaubitch, Eds., *Speech Dereverberation*, Signals and Communication Technology. Springer, 2010.

[6] H. Buchner, R. Aichner, and W. Kellermann, "Blind source separation for convolutive mixtures: A unified treatment," in *Audio Signal Processing for Next-Generation Multimedia Communication Systems*, Y. Huang and J. Benesty, Eds., pp. 255–293. Kluwer Academic Publishers, Boston, Feb. 2004.

[7] Y. Zheng, K. Reindl, and W. Kellermann, "BSS for improved interference estimation for blind speech signal extraction with two microphones," in *Proc. CAMSAP*, Dec. 2009.

[8] K. Reindl, Y. Zheng, A. Lombard, A. Schwarz, and W. Kellermann, "An acoustic front-end for interactive TV incorporating multichannel acoustic echo cancellation and blind signal extraction," in *Proc. 44th Asilomar Conference on Signals, Systems, and Computers*, Pacific Grove, CA, USA, Nov. 2010.

[9] R. Maas, A. Schwarz, Y. Zheng, K. Reindl, S. Meier, A. Sehr, and W. Kellermann, "A two-channel acoustic front-end for robust automatic speech recognition in noisy and reverberant environments," in *Proc. International Workshop on Machine Listening in Multisource Environments (CHiME 2011)*, Sep. 2011.

[10] Y. Takahashi, T. Takatani, K. Osako, H. Saruwatari, and K. Shikano, "Blind spatial subtraction array for speech enhancement in noisy environment," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 17, no. 4, pp. 650–664, May 2009.

[11] J. Even, H. Saruwatari, K. Shikano, and T. Takata, "Blind signal extraction based joint suppression of diffuse background noise and late reverberation," in *Proc. EUSIPCO*, Aug. 2010.

[12] L.C. Parra and C.V. Alvino, "Geometric source separation: merging convolutive source separation with geometric beamforming," *Speech and Audio Processing, IEEE Transactions on*, vol. 10, no. 6, pp. 352–362, Sep. 2002.

[13] A. Sehr, E.A.P. Habets, R. Maas, and W. Kellermann, "Towards a better understanding of the effect of reverberation on speech recognition performance," in *Proc. IWAENC*, Aug. 2010.

[14] E. Haensler and G. Schmidt, *Acoustic Echo and Noise Control: A Practical Approach*, Wiley-Interscience, 2004.

[15] M. Jeub and P. Vary, "Binaural dereverberation based on a dual-channel Wiener filter with optimized noise field coherence," in *Proc. ICASSP*, Mar. 2010.

[16] H. Kuttruff, *Room Acoustics*, Taylor & Francis, London, 2000.

[17] M. Cooke, J. Barker, S. Cunningham, and Xu Shao, "An audio-visual corpus for speech perception and automatic speech recognition," *The Journal of the Acoustical Society of America*, vol. 120, no. 5, pp. 2421–2424, 2006.

[18] D. Huggins-Daines, M. Kumar, A. Chan, A. W. Black, M. Ravishankar, and A. I. Rudnicky, "Pocketsphinx: A free, real-time continuous speech recognition system for hand-held devices," in *Proc. ICASSP*, May 2006.

[19] H. Christensen, J. Barker, and P. Green, "The CHiME corpus: a resource and a challenge for Computational Hearing in Multisource Environments," in *Proc. Interspeech*, 2010.