Acoustic Echo Control for Humanoid Robots

Adel El-Rayyes, Heinrich W. Löllmann, Christian Hofmann, Walter Kellermann
Multimedia Communications and Signal Processing, Friedrich-Alexander University Erlangen-Nürnberg
{*adel.el-rayyes,heinrich.loellmann,christian.hofmann,walter.kellermann*}*@fau.de*

Introduction

The design of acoustic signal processing algorithms for humanoid robots is a rather challenging task, e.g., [1]. For human-robot communication, the robot has to localize and track the active human speaker, and to understand the content of the conversation even in noisy and reverberant environments. In addition, the robot has to cope with ego-noise and acoustic feedback between the loudspeakers and microphones mounted in its head. In this contribution, a multi-channel system for Acoustic Echo Cancellation (AEC) and Beamforming (BF) is investigated for this purpose. The considered system aims at extracting the desired speaker in a noisy environment and suppresses the acoustic feedback between loudspeakers and microphones at the same time.

When designing a system for human-robot interaction, AEC is vital for robust speech interaction allowing for 'barge-in', but has gained only little attention so far [2, 3]. In addition, finding the optimal microphone positions for a BF scheme is important, as shown in [4]. Still, optimized designs may imply loudspeakers very close to microphones. However, a close loudspeaker-microphone distance may lead to clipping artefacts and may thus call for non-linear AEC schemes. To avoid this, careful sensor placement and an adaptive gain control are necessary. Furthermore, politeness, as it may be assumed in natural human-human interaction, cannot be assumed, leading to an increased amount of double-talk situations. These situations complicate the task of AEC, as most AEC systems require time intervals with dominant far-end signal in the microphone signals for identification of the echo paths. Therefore, an approach is proposed in this work, which jointly minimizes echo and interference and is rather robust to double-talk at the same time.

The remainder of this article is structured as follows: In the next section, a brief review of possible combinations of AEC and BF is given. Subsequently, the most suitable system and its individual components are discussed. Afterwards, an evaluation in terms of Echo Return Loss Enhancement (ERLE), Interference Reduction (IR) and Word Error Rate (WER) is carried out, followed by a conclusion and an outlook to future work in the last section.

The research leading to these results has received funding from the European Union's Seventh Framework Programme (FP7/2007-2013) under grant agreement n° 609465. It has been conducted as part of the project EARS (Embodied Audition for RobotS).

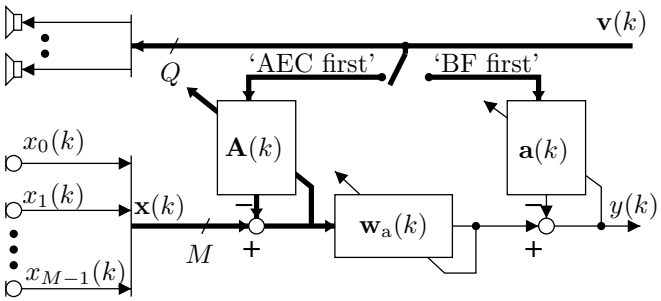


Figure 1: 'AEC first' or 'BF first' adapted from [5].

Combination of Beamforming and Acoustic Echo Cancellation

When considering joint adaptive BF and AEC, two basic configurations for AEC and BF can be applied, termed either as 'AEC first' or 'BF first' scheme [5], as illustrated in Figure 1. The Q loudspeaker and M microphone signals in Figure 1 are stacked in vectors $\mathbf{v}(k) = [\mathbf{v}_0^\mathrm{T}(k), \ldots, \mathbf{v}_q^\mathrm{T}(k), \ldots, \mathbf{v}_{Q-1}^\mathrm{T}(k)]^\mathrm{T}$ and $\mathbf{x}(k) = [\mathbf{x}_0^\mathrm{T}(k), \ldots, \mathbf{x}_m^\mathrm{T}(k), \ldots, \mathbf{x}_{M-1}^\mathrm{T}(k)]^\mathrm{T}$, respectively, where $\mathbf{v}_q = [v_q(k), \ldots, v_q(k - L_\mathrm{B} + 1)]^\mathrm{T}$, $\mathbf{x}_m = [x_m(k), \ldots, x_m(k - L_\mathrm{A} + 1)]^\mathrm{T}$, $m \in \{0, \ldots, M-1\}$ and $q \in \{0, \ldots, Q-1\}$. L_A and L_B denote the lengths of the employed Finite Impulse Response (FIR) AEC and BF filters, respectively. In addition, $\mathbf{A}(k)$ and $\mathbf{a}(k)$ denote the echo cancellation filters for the 'AEC first' and 'BF first' scheme, respectively. Furthermore, $\mathbf{w}_\mathrm{a}(k)$ represents the adaptive BF unit.

The 'AEC first' scheme offers a superior echo cancellation, in terms of echo reduction and coping with variations of the acoustic echo path at the price of a high computational complexity and the need for a sophisticated adaptation control. The 'BF first' scheme is less computationally demanding, since fewer echo paths have to be identified after the BF. However, the echo paths to be identified then incorporate the adaptive BF stage, which implies that the subsequent AEC has to model the time-variance of the adaptive beamformer with the associated convergence challenges for the filter adaptation.

The Generalized Sidelobe and Acoustic Echo Canceller (GSAEC) structure [6] offers a compromise between the two aforementioned methods, by placing the echo cancellation unit in the fixed BF path of a Generalized Sidelobe Canceller (GSC) [7]. In this solution, the AEC unit still has to model the fixed BF path in addition to the acoustic echo path, but is not impaired by the time variance of the interference cancelling path. However, in the GSAEC concept, the AEC unit does not profit from a converged

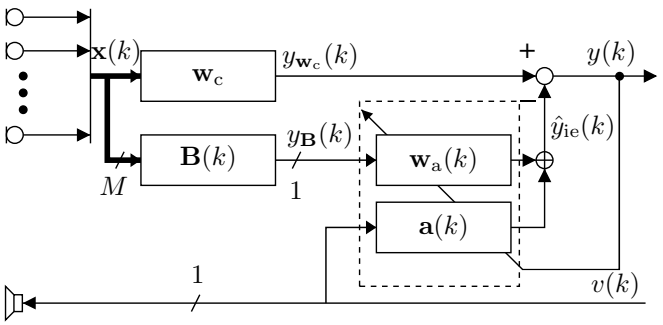


Figure 2: Generalized Echo and Interference Canceller, for mono AEC, adapted from [8].

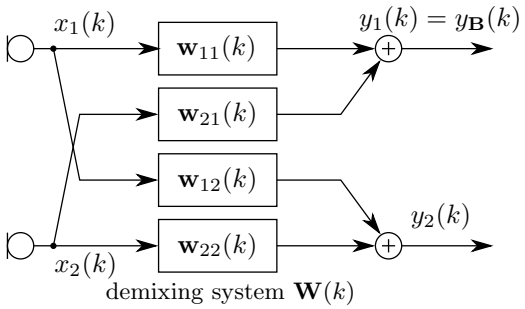


Figure 3: TRINICON-based blocking matrix.

Interference Canceller (IC) unit of the GSC structure. Therefore, a method is proposed in [8] which places the AEC unit in parallel to the IC unit, as illustrated in Figure 2. This structure offers a joint minimization of echoes and interference, leading to the so-called Generalized Echo and Interference Canceller (GEIC). In the following we consider how the individual components are modified for the robot audition scenario.

Components of GEIC structure

Fixed Beamformer $\mathbf{w}_\mathrm{c}(k)$: The fixed BF unit steers a time-invariant beam into the desired look direction ϕ_d. The beamformer output

$$y_{\mathbf{w}_\mathrm{c}}(k) = \mathbf{w}_\mathrm{c}^\mathrm{T}\mathbf{x}(k), \tag{1}$$

with fixed FIR filter coefficients

$$\mathbf{w}_\mathrm{c} = [\mathbf{w}_0^\mathrm{T}, \mathbf{w}_1^\mathrm{T}, \ldots, \mathbf{w}_{M-1}^\mathrm{T}]^\mathrm{T}, \tag{2}$$

aims at preserving signal components impinging on the microphone array from angular direction ϕ_d and suppressing components from other directions. Due to the limited sidelobe attenuation of beamformers, the beamformer output signal,

$$y_{\mathbf{w}_\mathrm{c}}(k) = y_\mathrm{d}(k) + y_\mathrm{e}(k) + y_\mathrm{i}(k), \tag{3}$$

contains, in addition to the desired signal component $y_\mathrm{d}(k)$, residual echo $y_\mathrm{e}(k)$ and interference $y_\mathrm{i}(k)$. As the microphones in a robot audition scenario are usually placed somewhere on the robot head, the assumption of a free-field sound propagation is not appropriate, and therefore, a common delay-and-sum beamformer is not suitable for the considered application. To account for the shadowing and scattering effect of the robot head, the BF weights, $\mathbf{w}_m$ with $m \in \{0, \ldots, M-1\}$, are estimated as relative impulse responses to a reference microphone for the desired look direction ϕ_d as proposed in [9].

Blocking Matrix $\mathbf{B}(k)$: As the fixed BF unit will not be able to suppress interfering noise sources completely, the Blocking Matrix (BM) is designed to retrieve estimates of the residual interference $y_\mathrm{i}(k)$ in the fixed beamformer output $y_{\mathbf{w}_\mathrm{c}}(k)$. In the proposed structure, the BM is based on the TRIple-N-Independent component analysis for CONvolutive mixtures (TRINICON) method [10] for Blind Source Separation (BSS). This algorithm aims at producing statistically independent output signals and thus allows the demixing of simultaneously recorded signals without the need of an explicit adaptation control. As TRINICON, in its conventional formulation, is constrained to demix only as many sources as microphone signals provided. In order to circumvent this drawback, a geometrically constrained version of TRINICON was introduced in [11] which exploits Direction of Arrival (DoA) information of a target source to suppress the target in one output channel and thereby separate all other signals from the target, representing the desired noise and interference reference at the output of the BM. In Figure 3, a TRINICON-based demixing system is depicted for $M_\mathrm{b} = 2$ microphone signals, $x_1(k)$ and $x_2(k)$, which are chosen from the available M microphone signals in Figure 2. The chosen microphones should preferably have a free line of sight to the desired source. Subsequently, only M_b microphone signals are used to estimate $P = 2$ statistically independent output signals, $y_1(k)$ and $y_2(k)$. This is achieved by convolving the microphone signals with the filter weights $\mathbf{w}_{p,m_\mathrm{b}}(k)$ of length L_b with $p \in \{1, \ldots, P\}$ and $m_\mathrm{b} \in \{1, \ldots, M_\mathrm{b}\}$, as shown in Figure 3. These filter weights form the demixing matrix $[\mathbf{W}(k)]_{p,m_\mathrm{b}} = \mathbf{w}_{p,m_\mathrm{b}}(k)$ which is updated using a gradient descent-based adaptation, which minimizes the joint cost function

$$J_\mathrm{total}(\mathbf{W}) = J_\mathrm{TRINICON}(\mathbf{W}) + \gamma_\mathrm{C} J_\mathrm{C}(\mathbf{W}) \tag{4}$$

of TRINICON $J_\mathrm{TRINICON}(\mathbf{W})$ and the geometric constraint $J_\mathrm{C}(\mathbf{W})$, respectively. The parameter γ_C in (4) can be adjusted according to the necessity of enforcing the geometric constraint. This system offers the possibility of producing a joint estimate of all interfering sources by suppressing the target source in one output signal. This is indicated in Figure 3 by $y_1(k) = y_\mathbf{B}(k)$, where the desired source is suppressed in $y_1(k)$. Thereby, effectively resulting in a $2L_\mathrm{b} \times 1$ blocking vector, which is represented by the first column of $\mathbf{W}(k)$. The other output signal $y_2(k)$, therefore, offers an estimate of the desired target signal, which can be exploited for an adaptation control in the subsequent echo and interference cancelling unit discussed in the following.

Echo and interference canceller $\mathbf{w}(k)$: The Echo and Interference Cancellation (EIC) stage exploits the available reference signals, i.e., the interference estimate $y_\mathbf{B}(k)$, produced by the BSS-based BM, and the loudspeaker signal $v(k)$, for estimating the residual echo and

interference in the fixed beamfomer output. The stacked weight vector

$$\mathbf{w}(k) = [\mathbf{w}_{\mathrm{a}}^{\mathrm{T}}(k), \mathbf{a}^{\mathrm{T}}(k)]^{\mathrm{T}} \tag{5}$$

of the EIC stage in Figure 2 is adapted using the GEIC output signal $y(k)$ and the vectors $\mathbf{y_B}(k)$ and $\mathbf{v}(k)$, containing the L_{EIC} most recent samples of the interference estimate and the loudspeaker signal, respectively. The update is performed according to

$$\mathbf{w}(k+1) = \mathbf{w}(k) + \Delta\mathbf{w}(k), \tag{6}$$

where the correction term $\Delta\mathbf{w}(k) = [\Delta\mathbf{w}_{\mathrm{a}}^{\mathrm{T}}(k), \Delta\mathbf{a}^{\mathrm{T}}(k)]^{\mathrm{T}}$ for the stacked weight vector can be estimated by a supervised adaptation algorithm. Here, for the adaptation a multi-channel frequency domain adaptive filter was used, e.g., [12]. By means of the adapted FIR filters, $\mathbf{w}_{\mathrm{a}}(k)$ and $\mathbf{a}(k)$, an echo and interference estimate, $\hat{y}_{\mathrm{i}}(k)$ and $\hat{y}_{\mathrm{e}}(k)$, respectively, are retrieved

$$\hat{y}_{\mathrm{ie}}(k) = [\mathbf{y}_{\mathbf{B}}^{\mathrm{T}}(k), \mathbf{v}^{\mathrm{T}}(k)]\mathbf{w}(k) \tag{7}$$

$$= \hat{y}_{\mathrm{i}}(k) + \hat{y}_{\mathrm{e}}(k). \tag{8}$$

This estimate is in turn subtracted from the fixed beamformer output signal, $y_{\mathbf{w}_{\mathrm{C}}}(k)$, to obtain a denoised source signal.

Experimental Setup and Evaluation Measures

For experimental evaluation, recordings were performed at a sampling rate of $f_{\mathrm{s}} = 16\,\mathrm{kHz}$ with the internal microphones of a NAO$^{\mathrm{TM}}$ v5 robot, which features $M = 4$ microphones mounted in its head, labeled A,...,D in Figure 4. The loudspeakers are visible in Figure 4 to the right and left of the head, illustrating the problem of close-coupled microphones and loudspeakers. Both loudspeakers can be assumed to play back an identical signal, therefore, only one echo path for each microphone needs to be considered. The robot was placed in a low-reverberant chamber with a reverberation time of $T_{60} \approx 50\,\mathrm{ms}$. Only one interferer and one target signal were active concurrently in addition to signals emitted from the robot's head-mounted loudspeakers. The desired source position was kept fixed throughout the recordings at $\phi_{\mathrm{d}} = 0°$ at a distance of 1m, i.e., the robot head was facing the desired source. The choice of the look-direction is based on the assumption that a robot would be facing its human interaction partner (assuming that the source localization and tracking system works perfectly). The interferer was positioned at $\phi_{\mathrm{i}} = \pm 40°$ at a distance of 1 m. In addition, the robot head was located at a height of 1.20 m and the loudspeakers, emitting the desired and interfering sources, were at a height of 1.45 m, leading to an elevation angle of $\theta \approx 15°$. The choice of θ is based on the fact that the robot is much smaller than a human, but would still tilt his head upwards, trying to face a human interaction partner, thereby decreasing θ approximately to the mentioned value. The loudspeaker at the desired look-direction emitted a speech signal of about 6 min duration

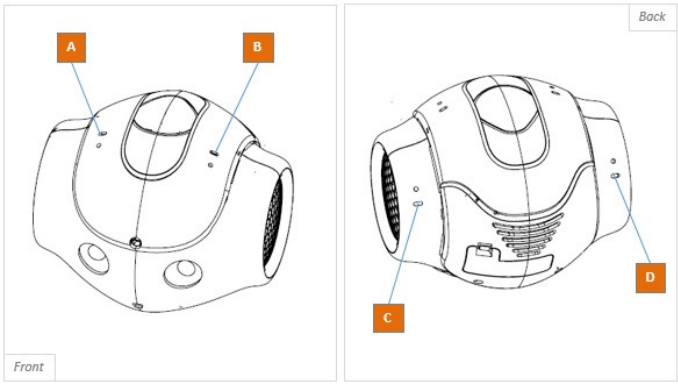


Figure 4: Drawing of the NAO v5$^{\mathrm{TM}}$ robot head, from [13].

(containing 200 utterances) taken from the GRID corpus [14]. The loudspeakers at the direction of the interfering source played back a female speech sequence with the same signal energy as the desired source, i.e., the Signal-to-Interference Ratio (SIR) was 0 dB. In addition, the loudspeaker signal was a male speech sequence with a higher signal energy than the desired source, leading to a negative near-end to far-end signal power ratio of −8 dB at the microphone position A, which only was active from 1:40 min onward. For the evaluation of the speech recognition performance, which is essential for robot audition, the Automated Speech Recognition System (ASR) engine PocketSphinx was used [15]. It employs a Hidden Markov Model (HMM)-Gaussian Mixture Model (GMM)-based acoustic model trained on clean speech from the GRID corpus [14], using MFCC+Δ+ΔΔ features (39 features in total) and cepstral mean normalization.

Furthermore, the signal processing chain was parametrized with the following values:

- length of fixed BF filters: $L_{\mathrm{f}} = 1024$
- length of BM filters: $L_{\mathrm{b}} = 1024$
- weighting of geometric constraint for TRINICON: $\gamma_{\mathrm{C}} = 0.5$
- length of EIC filters: $L_{\mathrm{EIC}} = 1024$.

The choice of L_{EIC} should allow for equal convergence behavior for the interference and echo cancelling path.

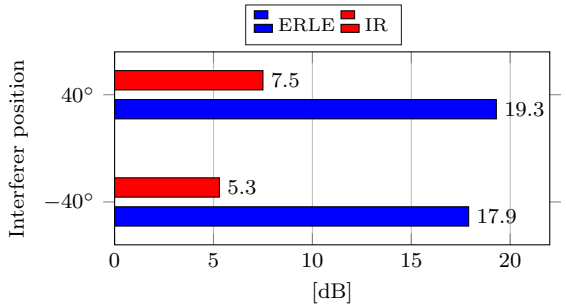
For the assessment of the echo and interference reduction, the *additional* ERLE and IR after the fixed beamformer was calculated, respectively, for the entire signal duration. Accordingly, the gains in ERLE and IR are estimated as

$$\mathrm{ERLE} = 10\log_{10}\left\{\frac{\mathcal{E}\{(y_{\mathrm{e}}(k))^2\}}{\mathcal{E}\{(y_{\mathrm{e}}(k) - \hat{y}_{\mathrm{e}}(k))^2\}}\right\}\,[\mathrm{dB}] \tag{9}$$

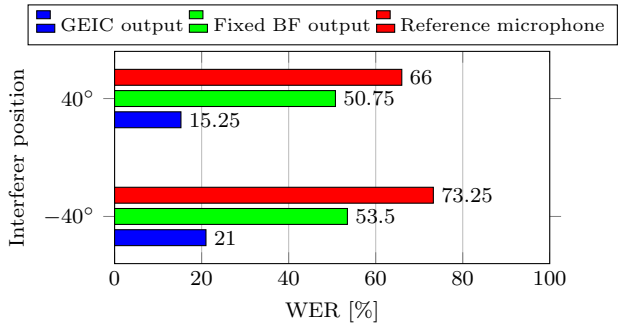
and

$$\mathrm{IR} = 10\log_{10}\left\{\frac{\mathcal{E}\{(y_{\mathrm{i}}(k))^2\}}{\mathcal{E}\{(y_{\mathrm{i}}(k) - \hat{y}_{\mathrm{i}}(k))^2\}}\right\}\,[\mathrm{dB}]. \tag{10}$$

In addition, the WER is evaluated as for the CHiME challenge [16]. The WER is the key evaluation metric, as it is essentially more important for robot audition that



(a) Average gain in ERLE and IR.



(b) Average WER.

Figure 5: Gains in ERLE and IR, and WER for two different interferer positions.

the robot understands its human interaction partner than achieving high ERLE and IR, whereas of course the latter are correlated with the former.

Results

Figure 5(a) shows that the gain in IR is only moderate compared to the gain in ERLE. This is due to the imperfect interference estimate delivered by the BM, which results from the fact that the level of the acoustic echo is significantly higher than the level of the distant interferer. The acoustic echo cancellation furthermore benefits from the fact that it can rely on a perfect reference signal. The generally low gain in ERLE and IR can be attributed to the frequent double-talk situations as target, interferer and loudspeaker are mostly simultaneously active.

In Figure 5(b), the results for the ASR engine are depicted in terms of the WER. While very high WERs are obtained when directly applying the ASR to an arbitrarily chosen reference microphone, some improvement is visible after processing by the fixed beamformer. This amounts to an improvement of roughly 20 percentage points in terms of WER. However, after substantial removal of echo and interference in the EIC stage, a significant reduction of the WER is obtained, which amounts to a dramatic improvement by more than 50 % absolute WER.

Conclusion

In this contribution, the GEIC structure has been modified to employ a BSS-based BM and applied for the task, of robot audition. The treated GEIC system is well suited for this task, due to its ability to efficiently suppress strong echoes caused by the short distance of the head loudspeakers and microphones, which results here in a significant reduction of the WER. Future work should include evaluation for more reverberant environment and if necessary employ additional echo and noise post filtering as well as dereverberation to enhance the ASR performance.

References

[1] Löllmann, H. W., Barfuss, H., Meier, S., Deleforge, A., and Kellermann, W.: Challenges in Acoustic Signal Enhancement for Human-Robot Communication, Proc. ITG Conf. Speech Communication, 2014.

[2] Takeda, R., Nakadai, K., Takahashi, T., Komatani, K., Ogata, T., and Okuno, H. G.: ICA-based efficient blind dereverberation and echo cancellation method for barge-in-able robot audition, IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP), 2009.

[3] Beh, J., Lee, T., Lee, I., Kim, H., Ahn, S., and Ko, H.: Combining acoustic echo cancellation and adaptive beamforming for achieving robust speech interface in mobile robot, IEEE/RSJ Int. Conf. Intelligent Robots, Systems (IROS), 2008.

[4] Tourbabin, V., Agmon, M., Rafaely, B., and Tabrikian, J.: Optimal real-weighted beamforming with application to linear and spherical arrays, IEEE Trans. Audio, Speech, Language Process., 20(9):2575-2585, 2012.

[5] Kellermann, W.: Acoustic echo cancellation for beamforming microphone arrays, In Microphone Arrays (pp. 281-306), Springer Berlin Heidelberg, 2001.

[6] Herbordt, W., and Kellermann, W.: GSAEC - Acoustic Echo Cancellation Embedded into the Generalized Sidelobe Canceller, Proc. European Signal Processing Conference (EUSIPCO), 2000.

[7] Griffiths, L., and Jim, C.: An alternative approach to linearly constrained adaptive beamforming. IEEE Trans. Antennas Propag., 30(1):27-34, 1982.

[8] Herbordt, W., Kellermann, W., and Nakamura, S.: Joint Optimization of LCMV Beamforming and Acoustic Echo Cancellation, Proc. European Signal Processing Conference (EUSIPCO), 2004.

[9] Gannot, S., Burshtein, D., and Weinstein, E.: Signal enhancement using beamforming and nonstationarity with applications to speech, IEEE Trans. Audio, Speech, Language Process., 49(8):1614-1626, 2001.

[10] Buchner, H., Aichner, R., and Kellermann, W.: Blind Source Separation for Convolutive Mixtures Exploiting Nongaussianity, Nonwhiteness, and Nonstationarity, Int. Workshop Acoustic Echo, Noise Control (IWAENC), 2003.

[11] Zheng, Y., Reindl, K., and Kellermann, W.: BSS for Improved Interference Estimation for Blind Speech Signal Extraction with Two Microphones, IEEE Int. Workshop Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP), 2009.

[12] Buchner, H., Herbordt, W., and Kellermann,W.: An efficient combination of multi-channel acoustic echo cancellation with a beamforming microphone array, Int. Workshop Hands-Free Speech Communication (HSC), 2001.

[13] NAOqi Documentation v2.1, URL: `http://doc.aldebaran.com/2-1/family/robots/microphone_robot.html`, retrieved 31.3.2016.

[14] Cooke, M., Barker, J., Cunningham, S., and Shao, X.: An audio-visual corpus for speech perception and automatic speech recognition, J. Acoust. Soc. Am., 120(5):2421-2424, 2006.

[15] Huggins-Daines, D., Kumar, M., Chan, A., Black, A.W., Ravishankar, M., and Rudnicky, A.I.: PocketSphinx: A free, real-time continuous speech recognition system for hand-held devices, IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP), 2006.

[16] Christensen, H., Barker, J., Ma, N., and Green, P.D.: The CHiME corpus: A resource and a challenge for computational hearing in multisource environments, INTERSPEECH, Sept. 2010.