INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

Robustness over time-varying channels in DNN-HMM ASR based human-robot interaction

José Novoa[1], *Jorge Wuth*[1], *Juan Pablo Escudero*[1], *Josué Fredes*[1], *Rodrigo Mahu*[1], *Richard Stern*[2], *Nestor Becerra Yoma*[1]
[1]Speech Processing and Transmission Lab., Universidad de Chile
[2]Robust Speech Recognition Group, ECE Dept., Carnegie Mellon University
nbecerra@ing.uchile.cl

Abstract

This paper addresses the problem of time-varying channels in speech-recognition-based human-robot interaction using Locally-Normalized Filter-Bank features (LNFB), and training strategies that compensate for microphone response and room acoustics. Testing utterances were generated by re-recording the Aurora-4 testing database using a PR2 mobile robot, equipped with a Kinect audio interface while performing head rotations and movements toward and away from a fixed source. Three training conditions were evaluated called Clean, 1-IR and 33-IR. With Clean training, the DNN-HMM system was trained using the Aurora-4 clean training database. With 1-IR training, the same training data were convolved with an impulse response estimated at one meter from the source with no rotation of the robot head. With 33-IR training, the Aurora-4 training data were convolved with impulse responses estimated at one, two and three meters from the source and 11 angular positions of the robot head. The 33-IR training method produced reductions in WER greater than 50% when compared with Clean training using both LNFB and conventional Mel filterbank features. Nevertheless, LNFB features provided a WER 23% lower than MelFB using 33-IR training. The use of 33-IR training and LNFB features reduced WER by 64% compared to Clean training and MelFB features.

Index Terms: speech recognition, human-computer interaction, time varying channels, locally-normalized filter banks.

1. Introduction

Robustness in real environments is one of the key features that must be provided by robot audition for successful human-robot interaction. Several challenges must be addressed in such environments, including but not limited to channel variability due to relative motion between the sound sources and the robot's microphones [1]. Despite the importance of this issue, little has been done toward this end. Recently, the second "CHiME" speech separation and recognition challenge [2] has defined a task to be addressed that includes movements of the speaker. However, they are limited to small head movements and the datasets were generated by convolving clean utterances with time-varying Binaural Room Impulse Responses (BRIRs) that mimic those movements. Baseline results for the ASR system used show that speaker motion has an impact over the keyword accuracy, degrading it by 4% on average with clean training, and improving it by 2% on average with reverberant training. Available speech databases that include relative motion between speaker and microphone are reviewed in [3] and a method of collecting large amounts of realistic noisy speech recordings with mobile robots was proposed. In the review, the databases that contain recordings with movements are limited to head movements only. In [1] several methods were proposed to improve the performance of robot audition in dynamically-changing acoustic environment, *i.e.*, moving sound sources, environmental changes caused by robot's motions, dynamical changes of the number of sound sources, and intrinsic variations in the sound sources themselves. While the study includes changes of the speaker's position between utterances, it does not consider data in which a speaker is moving within an utterance.

In scenarios where ASR is performed by moving robots, the corruption of speech by the additive noise of the robot's motors must be taken into account. In [4], the problem of motor noise in robot movement is addressed, and the authors introduced a new method for improving ASR in the presence of robot motor noise. The method is based on multi-condition training, maximum-likelihood linear regression (MLLR), and missing feature theory (MFT). In [5-6] egonoise suppression algorithms were employed. Also, sound localization and beam forming methods exploiting microphone arrays can be used to address the problem of additive noise in mobile robotics with multiple sound sources [7].

Distant speech recognition is also a challenge that has to be addressed in real environment and leads to degradation of speech recognition systems due to the effects of reverberation. Several authors have performed or proposed distant speech database recording for developing and testing techniques in real data. The development of an application for recording distant speech database, along with the actual recording using Kinect is described in [8]. Different techniques have been proposed or adapted to address the reverberation effect in automatic speech recognition (ASR): Wavelet Thresholding [9], Spectral Subtraction [10], Wiener Filtering [11], among others. In [12] a method in optimizing the speech enhancement techniques mentioned above is described. The optimization is specifically design to improve ASR. For evaluating the proposed method, the authors recorded real reverberant data in a multi-party conversation at different reverberation times in the Japanese and English languages. Word accuracy results show an improvement as great as 12% for a reverberation time of 80 ms, and 224% for a reverberation time of 940 ms, when compared to the results with no enhancement techniques.

Reference [13] describes the recording of a distant speech database recording in a multiroom smart home for evaluating combinations of techniques that operates at independent levels of the speech processing. The database is recorded in 8 channels and the nearest microphone is 2 meters away from the speaker. The utterances correspond to predefined expressions and phrase readings. The fusion of three enhancement techniques was evaluated: Beamforming, a technique that operates at the acoustic level [14]; the Driven Decoding Algorithm (DDA) [15], which operates at the decoding level by taking advantage of the availability of auxiliary transcripts; and ROVER

http://dx.doi.org/10.21437/Interspeech.2017-1308

(Recognizer Output Voting Error Reduction) [16] which operates at the ASR combination level, that is expected to improve the recognition results by providing the best agreement between the most reliable sources. The word error rate (WER) is reduced from 18.3% to 8.8%. Acoustic vectors were composed of 12 PLP (Perceptual Linear Predictive) coefficients, the energy, and the first and second order derivatives of these 13 parameters.

Since distant microphones are sensitive to speakers' body movements, the effects of distant speech and movement between the speaker and microphone produces a further degradation of ASR performance. The recognition rates are better when the speaker sits or moves only a little [17].

The present authors developed a novel set of speech features called Locally-Normalized Cepstral Coefficients (LNCC), which were initially proposed for robust Speaker Verification (SV) [18] and ASR [19]. LNCC features were inspired by Seneff's Generalized Synchrony Detector (GSD) [20] which performs a local normalization in the frequency domain in each auditory channel. LNFB features are LNCC features before the final DCT computation. The local normalization is achieved in the filter-bank space by dividing the output of a triangular frequency-weighting filter (which is similar to the triangular filter in conventional MFCC coefficients) by the output of a second frequency-weighting filter [18]. This normalization removes very coarse variations in the spectral shape that can be considered constant within both filters, such as overall tilt, which we assume arise mostly from channel variability. We refer to these two filters as the "numerator filter" and the "denominator filter," and their shape is an approximation to the frequency response of the numerator and denominator of the Seneff GSD operator:

$$Num_m(f)=\begin{cases}-\dfrac{2}{B}\left|f-f_m^C\right| & ,if\ \left|f-f_m^C\right|\leq\dfrac{B}{2}\\ 0 & ,otherwise\end{cases} \tag{1}$$

$$Den_m(f)=\begin{cases}\dfrac{2}{B}(1-d_{\min})\left|f-f_m^C\right|+d_{\min} & ,if\ \left|f-f_m^C\right|\leq\dfrac{B}{2}\\ 0 & ,otherwise\end{cases} \tag{2}$$

where the frequency variable f is in the Bark scale [21]. The shapes of these filters are shown in Figure 1. Given a channel m with center frequency f_m^C and bandwidth B, the LNFB feature m is defined as the log of the locally-normalized energy for channel m, LN_m:

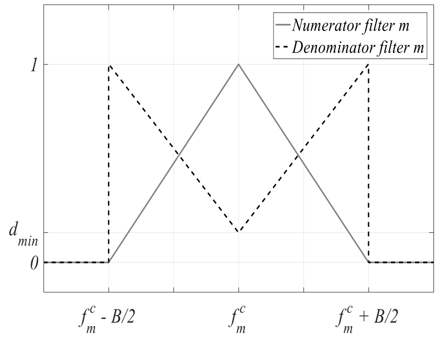

Figure 1: *Graphical representation of the m^th numerator filter (solid line) and the m^th denominator filter (dashed line).*

$$LNFB_m=\log(LN_m)=\log\left(LNNum_m/LNDen_m\right) \tag{3}$$

where $LNNum_m$ is the numerator filter energy, and $LNDen_m$ is the denominator filter energy. The parameter d_{min} prevents division by zero at the center frequency of each pair of numerator and denominator filters. One of the motivations behind the LNCC or LNFB features was to provide a set of parameters that were robust to time-varying channels such as those found in HRI environments. In these cases, temporal-trajectory filtering techniques, such as RASTA or CMN, are not applicable. In [21] LNCC was shown to reduce time-varying spectral tilt in a speaker verification task. In [19], the use of LNFB features provided significant reductions in WER in a DNN-HMM ASR system with channel mismatch.

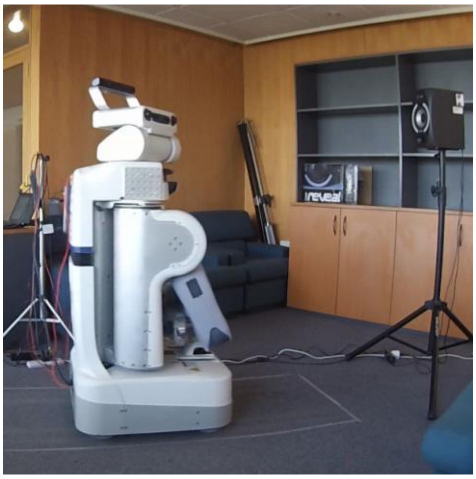

Figure 2: *PR2 robot employed in this study. The robot is equipped with a Microsoft Kinect that was used to record the database reproduced by the source that can be seen in left side of the picture.*

2. Robotic platform and database recording

Our experimental platform makes use of the PR2 (Personal Robot 2) shown in Figure 2. The PR2 is a state-of-the-art mobile manipulation robot equipped with a Microsoft Kinect among other sensors. We re-recorded the testing utterances of the Aurora-4 database in a meeting room (Figure 3) including different specifications of the relative motion between the robot and the speaker. The audio source used for reproducing the audios was a Monitor TANOY 501a. The recording was performed by the PR2's Kinect, which contains a four-microphone array. The channels of the four-channel audio recorded by the Kinect were summed, generating a single-channel speech signal.

The recording procedure considers the relative movements of the robot microphones with respect to the source by simultaneously applying translational movement to the robot body and angular rotation to the robot head upon which the Kinect is mounted.

2.1. Robot Displacement

Three velocities, v_1, v_2, and v_3, of displacement of the PR2 Robot from the Point P1 to the Point P3 shown in Figure 3 were

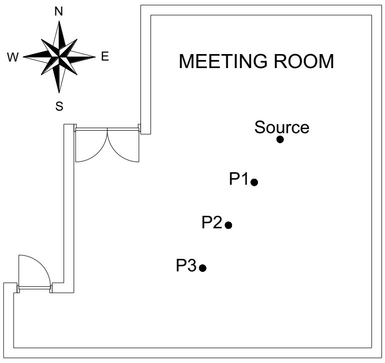

Figure 3: *PR2 Robot movement from Point P1 to Point P3 during the recording of the test sets. The sound source is located at the Source point. The route was made repeatedly during the utterances recording, moving in each cycle from Point P1 (located 1m from the source) to Point P3 (located 3m from the source) and back again. Recordings in the static condition were obtained at Point P1.*

established; 0.30 m/s, 0.45 m/s and 0.60 m/s, respectively. Those velocities were motivated by the discussions in [22], where a robot approached to a seated person at 0.2 m/s and 0.4 m/s. In that work, none of the human participants found those velocities too fast. In order to perform smooth movements that do not provoke some kind of rejection by humans, a speed factor function was established along with the robot movement. By doing so, the actual velocity of the robot is given by (4):

$$v_{smooth_i}(t) = \begin{cases} v_i \cdot t & ,if\ 0 < t < 1 \\ v_i & ,if\ 1 \le t < t_{d_i} - 1 \\ v_i \cdot (t_{d_i} - t) & ,if\ t_{d_i} - 1 \le t \le t_{d_i} \end{cases} \quad (4)$$

where v_{smooth_i} is the actual displacement velocity of the robot for each v_i, $i = 1,2,3$. Additionally, t_{d_i} is the time required to go from Point P1 to P3 when moving with velocity v_{smooth_i} described in (4), and can be computed as $t_{d_i} = 1 + d / v_i$, where d is the distance between P1 and P3. Additionally, a test database recording was performed with the robot remaining at Point P1. In this way, four displacement conditions were considered for the recordings: one with the robot remaining at P1, and three with the robot performing translational movements between P1 and P3 with velocities v_{smooth_1}, v_{smooth_2} and v_{smooth_3}.

2.2. Head movement

For each of the four displacement conditions described above, the robot makes turns with the head as shown in Figure 4(a). The head moves repeatedly from –150º to 150º and back at different angular velocities. The source is located at 0°. For the motion described above, three angular velocities ω_i for the robot head were set: 0.14 rad/s, 0.28 rad/s, and 0.42 rad/s. The selected angular velocities correspond to the angular speed of the head rotation necessary for the Robot to follow with the head a target located two meters away from it and moving with tangential velocities of 2 km/h, 3 km/h and 4 km/h, respectively, as shown in Figure 4(b). The fourth motion condition was zero, fixing the robot's head at 0° (*i.e.,* oriented towards the source) for each robot displacement described above. The combination of four conditions for robot displacement and four robot head movements produces 16 test database recording conditions.

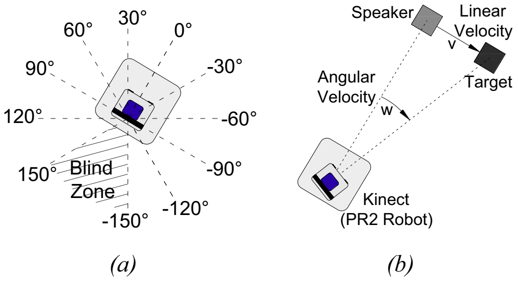

Figure 4: *(a) Movement of the PR2 robot head that was performed during the utterances recording. The head moves repeatedly from –150º to 150º and back at different angular velocities. Recordings with static head are performed at 0º. The source is located at 0°. (b) The selected angular velocities for the robot's head correspond to the speed of head rotation necessary for the Robot to follow with the head a target located two meters away and moving with linear velocities of 2 km/h, 3 km/h and 4 km/h, respectively. The source is located at 0°.*

3. Experiments

Speech recognition experiments were performed using the Kaldi Speech Recognition Toolkit [23]. Three training sets were employed, referred to as Clean, 1-IR, and 33-IR. The Clean training dataset consisted of the original utterances of the Aurora-4 database. The Clean training set consists of 7138 utterances from 83 speakers and contains only clean data recorded with a Sennheiser HMD 414 microphone.

For the 1-IR and 33-IR training data, the robot was turned off so that the recordings capture the characteristics of the room and the Kinect microphones, but not the effects of the additive noise produced by the motors of the robot. (We will address this issue in future work.) Using Farina's sine sweep method [24], an Impulse Response (IR) was computed with the robot placed at 1m from the source (Point P1) and with the Kinect microphones oriented toward the source. A 4-channel sweep was recorded and the individual channels were mixed as described above to generate a single channel sweep used to generate the IR. The 1-IR training set was generated by convolving the 7138 utterances from the clean training set of the Aurora-4 database with the estimated IR. Also using the Farina's sine sweep method [24], several IRs were computed with the robot placed at different distances from the source and with different orientations of the Kinect microphones with respect to the source. The robot was placed in Points P1, P2 and P3, located at 1m, 2m and 3m from the source, respectively (Figure 3). For each point, the head was oriented at 11 different angles with respect to the source. The head angle was varied from –150º to 150º in steps of 30º. Angle 0º corresponds to the Kinect microphones oriented towards the source. For each of the 33 configurations, a 4-channel sweep was recorded with the Kinect, and the individual channels were mixed as described above to generate the single-channel sweep used to generate each of the 33 IRs. For creating the 33-IR training set, 25% of the clean training set of the Aurora-4 database was convolved with the estimated IRs at 1m distance and angle 0º with respect to the source. The remaining 75% of the clean training set was convolved with the remaining 32 IRs in such a way that the 32 IRs were evenly distributed across the signals.

In this paper, we compare results obtained using the MelFB and LNFB feature vectors. The DNN-HMM system is composed of DNNs with seven hidden layers and 2048 units per layer each, using a context window of 11 frames. The DNN-HMM systems were trained using alignments from a GMM-HMM recognizer trained with the same data. The GMM-HMM systems were trained using MFCC features, linear discriminant analysis (LDA), and maximum likelihood linear transforms (MLLT), according to the tri2b Kaldi Aurora-4 recipe. First, a monophone system was trained; second, the alignments from that system were employed to generate an initial triphone system; and finally, the triphone alignments were employed to train the final triphone system. The number of units of the output DNN layer was equal to the number of Gaussians in the corresponding GMM-HMM system. The standard 5K lexicon and trigram language model were used.

4. Results and Discussion

Results were obtained for a total of 96 experimental conditions consisting of all permutations of the four displacement velocities v (0, 0.3, 0.45, and 0.6 m/s), four head angular velocities ω (0, 0.28, 0.42, 0.56 rad/s), two types of feature extraction procedures (MelFB and LNFB), and three sets of training data (Clean, 1-IR, and 33-IR). Table 1 describes the WER obtained for each experimental condition. The lowest WER for each testing condition is highlighted in bold. As can be seen in Table 1, the best results are observed for LNFB in all cases for each training condition. Note that 1-IR training achieves the best WER only for the case of a static robot, where test and training conditions match perfectly. Otherwise, the use of 33-IR training condition with LNFB features leads to a WER reduction greater than 54% when compared with a baseline system with MelFB features and Clean training.

On average, LNFB features outperform MelFB over all training conditions. The WER for LNFB is 19% (relative) less than for MelFB, in the Clean training condition, and 23% less in the 1-IR and 33-IR training conditions. A comparison of training conditions reveals that the use of 1-IR training leads to 35% and 32% WER reductions for LNFB and MelFB, respectively, compared to Clean training. This improvement most likely reflects the incorporation of the room and Kinect microphones responses in the training data. For the 33-IR training conditions, the WER is reduced by 56% and 53% with respect to Clean training, for LNFB and MelFB, respectively. These greater reductions in WER are due to additionally incorporating into the training data the three source-microphone distances and 11 head angles for each distance. In this way, the DNN-HMM system can also compensate for the channel variability caused by the robot movements.

As can be seen, the WER obtained when the robot is in motion is worse than when the robot is static at 1 m from the source. This degradation increases linearly with the displacement velocity, and can be as high as 202% and 253% for LNFB and MelFB, respectively, for the greatest velocity. We believe that part of the degradation is caused by the robot motors noise, which was found to increase linearly with velocity. The effect of channel variability given by the robot movement towards and away from the source also increases with the displacement velocity, leading to an additional degradation. It is worth mentioning that the use of LNFB features reduces the WER respect to the WER obtained with conventional MelFB, confirming the natural robustness of the LNFB features with channel variability and channel mismatch.

Table 1: *WERs obtained using MelFB and LNFB features with different training conditions and different velocities of robot displacement and head rotation.*

		Training Condition					
Testing Condition		**Clean**		**1-IR**		**33-IR**	
v **[m/s]**	ω **[rad/s]**	**MelFB**	**LNFB**	**MelFB**	**LNFB**	**MelFB**	**LNFB**
0	**0**	9.3	8.6	5.5	**5.4**	6.2	6.2
	0.28	52.4	43.8	29.0	22.2	14.8	**12.9**
	0.42	53.2	41.4	28.7	19.4	14.6	**11.8**
	0.56	54.5	42.1	28.1	19.9	14.3	**11.9**
0.3	**0**	36.2	25.9	18.4	12.9	15.9	**10.6**
	0.28	77.6	66.6	52.8	42.4	32.6	**27.5**
	0.42	77.0	65.8	52.4	43.9	33.2	**27.2**
	0.56	79.7	67.1	56.1	44.9	34.8	**27.5**
0.45	**0**	45.1	30.3	21.3	15.9	17.9	**12.3**
	0.28	83.3	68.6	62.3	49.2	42.2	**33.0**
	0.42	83.8	68.7	62.4	49.5	43.1	**33.0**
	0.56	84.4	70.5	65.5	49.8	43.5	**33.0**
0.6	**0**	55.3	33.4	28.5	19.5	26.5	**15.5**
	0.28	86.4	73.4	69.1	53.5	50.3	**37.5**
	0.42	85.8	69.4	66.6	50.8	48.5	**36.5**
	0.56	86.9	73.1	68.7	54.0	51.1	**39.5**

WER is also worse when the robot head is undergoing rotational motion compared to when it is static. Nevertheless, this degradation is relatively independent of angular velocity, and can be as high as 151% and 116% for LNFB and MelFB, respectively, for the greatest velocity. It is worth mentioning that the percentage of occluded frames in each testing condition, *i.e.* frames for which the path from the source to the Kinect microphones is blocked by the Kinect encasement, is the same for each head angular velocity, except for the static head condition which does not produce any occluded frames. Moreover, the noise power of the head motors was found to be independent of the head angular velocity, except for the static head condition which produces no head motor noise.

5. Conclusions

Locally-Normalized Filter-Bank features (LNFB) and DNN-HMM training strategies were employed to address the problem of time-varying channels in speech recognition based human-robot interaction. Time-varying channels were generated by performing displacement movements and head rotations at different speeds with respect to a source location that remained fixed. The use of 33-IR training produced reductions in WER greater than 50% compared to Clean training with both LNFB and MelFB. However, LNFB provided a WER 23% lower than MelFB with 33-IR. When compared with Clean training and MelFB, 33-IR and LNFB led to a reduction in WER equal to 64%. The reduction of the additive noise effect due to the robot engine and the application of beam forming methods are proposed for future research.

6. Acknowledgements

The research reported here was funded by Grant Conicyt-Fondecyt 1151306 and ONR Grant Nº N62909-17-1-2002. José Novoa was supported by Grant CONICYT-PCHA/Doctorado Nacional/2014-21140711.

7. References

[1] K. Nakadai, G. Ince, K. Nakamura, and H. Nakajima, "Robot audition for dynamic environments," in *Proceedings of IEEE International Conference on Signal Processing, Communication and Computing (ICSPCC)*, 2012, Hong Kong, China, pp. 125-130.

[2] E. Vincent, J. Barker, S. Watanabe, J. Le Roux, F. Nesta, and M. Matassoni, "The second 'CHiME' speech separation and recognition challenge: datasets, tasks and baselines," in *Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2013, Vancouver, BC, Canada, pp. 126-130.

[3] J. Le Roux, E. Vincent, J. R. Hershey, and D. P. Ellis, "MICbots: collecting large realistic datasets for speech and audio research using mobile robots," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, South Brisbane, Australia, pp. 5635-5639.

[4] Y. Nishimura, M. Nakano, K. Nakadai, H. Tsujino, and M. Ishizuka, "Speech recognition for a robot under its motor noises by selective application of missing feature theory and MLLR," in *Proceeding of ISCA Tutorial and Research Workshop on Statistical and Perceptual Audition*, Pittsburgh, PA, USA, pp. 53-58, 2006.

[5] G. Ince, K. Nakadai, T. Rodemann, Y. Hasegawa, H. Tsujino and J. i. Imura, "Ego noise suppression of a robot using template subtraction," in *Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2009, St. Louis, MO, USA, pp. 199-204.

[6] T. Tezuka, T. Yoshida and K. Nakadai, "Ego-motion noise suppression for robots based on semi-blind infinite non-negative matrix factorization," in *Proceedings of IEEE International Conference on Robotics and Automation (ICRA)*, 2014, Hong Kong, China, pp. 6293-6298.

[7] J. M. Valin, F. Michaud, and J. Rouat, "Robust 3D localization and tracking of sound sources using beamforming and particle filtering," in *Proceedings of International Conference on Audio, Speech and Signal Processing (ICASSP)*, 2006, Toulouse, France, pp. 221–224.

[8] D. Schnelle-Walka, S. Radeck-Arneth, C. Biemann, and S. Radomski, "An open source corpus and recording software for distant speech recognition with the Microsoft Kinect," in *Proceedings of 11. ITG Symposium on Speech Communication,* 2014, Erlangen, Germany, pp. 1-4.

[9] D. L. Donoho, "Denoising by soft thresholding," *IEEE Transactions on Information Theory,* vol.41, no.3, pp. 613-627, 1995.

[10] R. Gomez, J. Even, H. Saruwatari, and K. Shikano, "Fast dereverberation for hands-free speech recognition," in *Proceedings of IEEE Workshop HSCMA*, 2008, Trento, Italy, pp. 140-143.

[11] R. Gomez and T. Kawahara, "Optimizing spectral subtraction and Wiener filtering for robust speech recognition in reverberant and noisy conditions," in *Proceedings of International Conference on Audio, Speech and Signal Processing (ICASSP)*, 2010, Dallas, TX, USA, pp. 4566-4569.

[12] R. Gomez, T. Kawahara, K. Nakamura, and K. Nakadai, "Multi-party human-robot interaction with distant-talking speech recognition," in *Proceedings of the Seventh Annual ACM/IEEE International Conference on Human-Robot Interaction*, 2012, Boston, MA, USA, pp. 439-446.

[13] B. Lecouteux, M. Vacher, and F. Portet, "Distant speech recognition in a smart home: Comparison of several multisource ASRs in realistic conditions," in *Proceedings of INTERSPEECH*, 2011, Florence, Italy, pp. 2273-2276.

[14] X. Anguera, C. Wooters, and J. Hernando, "Acoustic beamforming for speaker diarization of meetings," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 7, pp. 2011–2022, 2007.

[15] B. Lecouteux, G. Linarès, J. Bonastre, and P. Nocéra, "Imperfect transcript driven speech recognition," in *Proceedings of INTERSPEECH*, 2006, Pittsburgh, PA, USA, pp. 1626–1629.

[16] J. G. Fiscus, "A post-processing system to yield reduced word error rates: Recognizer Output Voting Error Reduction (ROVER)," in *Proceedings of IEEE Workshop ASRU*, 1997, Santa Barbara, CA, USA, pp. 347-354.

[17] Y. Liu, P. Zhang, and T Hain, "Using neural network front-ends on far field multiple microphones based speech recognition," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2014, Florence, Italy, pp. 5542-5546.

[18] V. Poblete, F. Espic, S. King, R. M. Stern, F. Huenupán, J. Fredes, and N. Becerra Yoma, "A perceptually-motivated low-complexity instantaneous linear channel normalization technique applied to speaker verification," *Computer Speech & Language*, vol. 31, no. 1, pp. 1-27, 2015.

[19] J. Fredes, J. Novoa, S. King, R. M. Stern, and N. Becerra Yoma, "Locally-normalized filter banks applied to deep neural network-based robust speech recognition," *IEEE Signal Processing Letters*, vol. 24, no. 4, pp. 377-381, 2017.

[20] S. Seneff, "A joint synchrony/mean-rate model of auditory speech processing," *Journal of Phonetics*, vol. 16, pp. 55-76, 1988.

[21] J. Fredes, J. Novoa, S. King, R. M. Stern, and N. Becerra Yoma, "Robustness to additive noise of locally-normalized cepstral coefficients in speaker verification," in *Proceedings of INTERSPEECH*, 2015, Dresden, Germany, pp. 3011-3015.

[22] K. Dautenhahn, M. Walters, S. Woods, K. L. Koay, C. L. Nahaniv, E. A. Sisbot, R. Alami, and T. Siméon, "How may I serve you? A robot companion approaching a seated person in a helping context," in *Proceedings of ACM Conference on Human-Robot Interaction (HRI)*, 2006, Salt Lake City, UT, USA, pp. 172-179.

[23] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlíček, Y. Qian, P. Schwarz, J. Silovský, G. Stemmer, and K. Veselý, "The Kaldi speech recognition toolkit," in *Proceedings of ASRU*, 2011, Waikoloa, HI, USA, N° EPFL-CONF-192584.

[24] A. Farina, "Simultaneous measurement of impulse response and distortion with a swept sine technique," In *Proceedings of 108th AES Convention*, 2000, Paris, France, p. 5093.