Channel Selection for Distant Speech Recognition Exploiting Cepstral Distance

Cristina Guerrero[1,2], *Georgina Tryfou*[1,2], *Maurizio Omologo*[2]

[1]University of Trento, Trento, Italy
[2]Fondazione Bruno Kessler, Trento, Italy
guerrero@fbk.eu, tryfou@fbk.eu, omologo@fbk.eu

Abstract

In a multi-microphone distant speech recognition task, the redundancy of information that results from the availability of multiple instances of the same source signal can be exploited through channel selection. In this work, we propose the use of cepstral distance as a means of assessment of the available channels, in an informed and a blind fashion. In the informed approach the distances between the close-talk and all of the channels are calculated. In the blind method, the cepstral distances are computed using an estimated reference signal, assumed to represent the average distortion among the available channels. Furthermore, we propose a new evaluation methodology that better illustrates the strengths and weaknesses of a channel selection method, in comparison to the sole use of word error rate. The experimental results suggest that the proposed blind method successfully selects the least distorted channel, when sufficient room coverage is provided by the microphone network. As a result, improved recognition rates are obtained in a distant speech recognition task, both in a simulated and a real context.

Index Terms: distant speech recognition, channel selection, cepstral distance, reverberation

1. Introduction

State-of-the-art speech recognizers achieve highly accurate recognition rates when the speech signal is recorded by a close-talking microphone, e.g., a head-set. In a distant-talking setting, the distance between the speaker and the microphone introduces challenges to speech recognition as it must deal with interferences such as background noise, competing speakers, and reverberation. In order to address such difficulties, various approaches have been studied [1, 2, 3, 4, 5]. Many of the most effective solutions exploit multiple microphones, either as compact or distributed arrays, to capture multiple distorted instances of the same source signal. These multiple signals can be utilized at different stages of the recognition system, such as at front-end processing, decoding, or post-decoding stages.

Concerning the problem of distant speech recognition (DSR) in a multi-microphone setting, it is reasonable to categorize the various approaches present in the literature as solutions focused on the combination or the selection of elements, which can be for example: signals, features, or decoding outputs [1]. In this work, we aim our attention to signal processing approaches. The most representative solutions at this level are noise reduction, single and mutiple channel enhancement [6], beamforming [1, 7] and channel selection [8]. The application of beamforming on distantly located microphones is fundamentally restricted due to the effects introduced by spatial aliasing. In a setting in which the microphones of the network are largely distributed in a room, a valid alternative for combining different signals is given by channel selection (CS). The goal of CS is to identify, among all the available input signals, the one that leads to the best recognition accuracy. In a real application context this should work dynamically, selecting the optimal signal at each speech input. A scoring mechanism is required in order to identify the best channel.

The signal related scores explored for CS can be either informed or blindly computed. Informed scores assume the availability of some knowledge or reference information, and are often used to establish an upper-bound of performance. Such scores include those computed from room impulse responses [9], exploiting for example early to late reflections, signal to noise ratio (SNR) [10], and CS based on the position and orientation of the speaker [11]. Proposed blind scores include those computed from the energy of the signal, cross-correlation between signals [12], the variance of the energy envelope [8], and the modulation spectra of the original and of the beamformed signals [13]. According to an exhaustive survey on various CS approaches, presented in [14], the method based on the envelope of the signal energy achieved the highest recognition accuracy.

Distance measures for speech processing were introduced and applied primarily by the speech coding community to quantify the distortion introduced by the coding process [15, 16, 17]. Similar scoring strategies, although with different goals, are also exploited by other communities, such as speech enhancement and speech recognition. A lot of efforts have been made in order to produce scores that objectively evaluate the effectiveness of different speech processing techniques [18], as for example the cepstral distance (CD), the log-likelihood ratio, and the frequency-weighted segmental SNR. Such scores have been shown to correlate well with subjective evaluation of signal quality [19]. It is therefore reasonable to assume that the use of objective signal quality scores can lead to a meaningful selection of the least distorted channel, among the signals of a distributed microphone network. Particularly, the CD is long known for its effectiveness and flexibility in different application fields [20].

In this work, we propose the use of the CD as a scoring function for the selection of the best, or least distorted, channel in a real, multi-microphone setting. In a first case we assume the availability of the clean, or close-talk signal. In this case, we show that the use of cepstral distance can lead to a meaningful selection of the best channel. In a second case, the clean signal is no longer available. On the other hand, we derive a reference, in the log-magnitude spectrum domain, that can be used in order to assign a distance to each of the available channels, and facilitate the selection of the least distorted one. The exper-

imental results prove the success of the proposed blind channel selection in both simulation and real data.

The remainder of this paper is structured as follows. In Section 2, the details of the proposed scoring functions and CS methods are illustrated. Section 3 describes the experimental activities performed in order to assess the value of various CS methods. A discussion of the experimental findings is elaborated in Section 4. Finally, the conclusions of this work are presented in Section 5.

2. Cepstral distance based channel selection

In a multi-microphone scenario, with many microphones distributed in the room, it is reasonable to assume that an objective measure of signal quality would be advantageous in detecting the least distorted channel. Perhaps the most intuitive objective measure for signal quality, that applies well in cases of reverberation, is the CD. Cepstrum-based comparisons are equivalent to comparisons of the smoothed log spectra of the signals [20]. In this domain, the reverberation effect can be viewed as additive [21]. Furthermore, as discussed in [22], the CD has a particular frequency domain interpretation in terms of relationship between a set of signals and their geometric mean spectrum. Here, we study the use of CD for channel selection, in an informed and a blind fashion.

2.1. Informed channel selection

Assuming the availability of the close-talk signal, $x(t)$, and a multi-microphone setting, let

$$x_m(t) = x(t) * h_m(t) \tag{1}$$

be the signal captured by microphone m, where $h_m(t)$ is the related impulse response (IR). Here, $x_m(t)$ is not distorted by environmental noise. The CD between the close-talk reference and the distorted signal is defined as [19]:

$$d(\boldsymbol{c}_x, \boldsymbol{c}_m) = \frac{10}{log10}\sqrt{2\sum_{k=1}^{p}[\boldsymbol{c}_x(k) - \boldsymbol{c}_m(k)]^2} \quad , \tag{2}$$

where $\boldsymbol{c}_x$ and $\boldsymbol{c}_m$ are the cepstral coefficient vectors of the close-talk and distorted signals respectively, and p is the number of cepstral coefficients used.

From the set of CDs between the reference and all the available channels, the least distorted one can be selected as follows:

$$\hat{M}_x = \underset{m}{\operatorname{argmin}} \quad d(\boldsymbol{c}_x, \boldsymbol{c}_m). \tag{3}$$

2.2. Reference extraction for blind channel selection

In a real scenario, the close-talk signal is not available. Therefore, we propose a non-intrusive method for cepstral based channel selection, which exploits a multi-microphone distant speech recognition scenario for the estimation of a reference. When the speaker is oriented towards one of the many available distributed microphones, and/or is located at a distance lower than the critical distance [23], it is observed that for the corresponding signal, the direct component is generally stronger than the reverberated part. Other channels, whose energy is attenuated by the head of the speaker and other possible propagation effects, are expected to be more affected by reverberation. Based on this observation, we can average in the log-magnitude spectrum domain as follows:

$$\hat{R}(t,\omega) = \frac{1}{M}\sum_{m}\log|X_m(t,\omega)| \quad , \tag{4}$$

where $X_m(t,\omega)$ is the short-time Fourier transform (STFT) of the signal captured by microphone m, and M is the total number of microphones. This represents the corresponding geometric mean spectrum [22] and, using Eq. 1 this can be rewritten into:

$$\hat{R}(t,\omega) = \log|X(t,\omega)| + \frac{1}{M}\sum_{m}\log|H_m(t,\omega)| \quad , \tag{5}$$

where $X(t,\omega)$ and $H_m(t,\omega)$ are the STFT of the clean signal and m-th IR respectively. In Eq. 5, the first term is the log-magnitude spectrum of the close-talk signal, and the second term represents an estimation of the average reverberation of the room, based on the available microphone channels. Let us assume that one microphone signal is better than the others in terms of direct to reverberant ratio. The basic assumption is that such a signal will be characterized by a larger distance from the resulting geometric mean spectrum. Therefore, from the set of CDs between the geometric mean spectrum, $\hat{R}(t,\omega)$, and all the available channels, the least distorted one can be selected as follows:

$$\hat{M}_{\hat{R}} = \underset{m}{\operatorname{argmax}} \quad d(\boldsymbol{c}_{\hat{R}}, \boldsymbol{c}_m) \quad , \tag{6}$$

where $\boldsymbol{c}_{\hat{R}}$ and $\boldsymbol{c}_m$ are the cepstral coefficient vectors of the reference and of the microphone m, respectively.

So far, it was assumed that the only source of distortion of the M acquired signals was the reverberation. However, in a real setting this is rarely true, as environmental noise also exists. Given the purpose of this study, here we assume that the reverberation effect dominates over a background noise that affects all the microphones.

3. Experiments

3.1. Experimental setup

The experimental scenario is taken from the DIRHA Project setup [24], see http://dirha.fbk.eu. The experiments are performed within the livingroom, a room characterized by a T60 of about $0.75s$ and equipped with multiple largely-spaced microphones, located on the walls and ceiling. A subset of 6 microphones are selected for this study, as shown in Figure 1.

The training data consists of 7138 simulated reverberant utterances, derived from the full clean Wall Street Journal (WSJ0-5k) [25] training set. This training set was simulated using recorded IRs, which consider only channels in which the speaker position/orientation (POSORIs) is direct towards a microphone.

The test material is extracted from the WSJ0-5k sub-set of the DIRHA-English [26] corpus, which includes data recorded in the real livingroom. With regard to the test sets, in order to focus the analysis on the different CS methods, two scenarios are considered:

- In the first one, the speaker POSORI is always direct in respect to one microphone. Such a setting narrows the DSR problem, allowing us to perform an intuitive analysis of the correlation between signal distortion and recognition performance. For this scenario, simulated data is generated under two specific POSORI configurations (DirSim). Additionally, real data is extracted based on a set of 7 different POSORIs (DirReal). Figure 1 depicts the POSORIs used for the first scenario.
- The second scenario incorporates a set of 36 mixed POSORIs, for each of the simulated and real cases

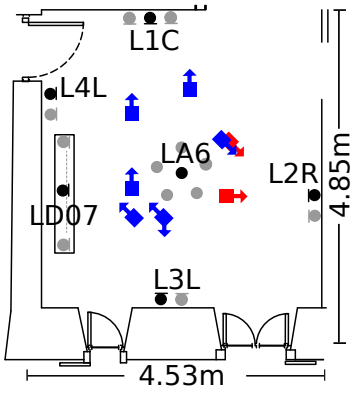

Figure 1: DIRHA Setting. Dots in black indicate the microphones used. Arrows show position/orientation of the speakers for the Direct scenarios, in red for simulated data, and in blue for real data.

(MixSim and MixReal). In this scenario, the adopted POSORIs are distributed in the room and are not only direct.

For all the simulated data, the close-talk signals were recorded in the FBK recording studio, while for the real data, these were captured by a head-set. The DirSim, MixSim and MixReal datasets are composed by 410 utterances each. DirReal dataset is composed by 82 utterances. An ideal voice activity detection is assumed to be applied over real data.

3.2. Channel selection methods

The following CS methods are included in the evaluation:

- **CDi** is the proposed informed CS method that uses the close-talk reference, as explained in Section 2.1.
- **CDref** is the blind proposed CD method that uses the geometric mean spectrum as a reference, as described in Section 2.2.
- **EV** [14] is the state-of-the-art CS method, based on Envelope Variance. It consists in selecting the channel as:

$$\hat{C} = \underset{m}{\text{argmax}} \sum_{k} \frac{V_m(k)}{\max_m(V_m(k))} \quad , \tag{7}$$

where $V_m(k)$ is a variance measure computed from the sub-band envelopes of the mean-subtracted filter-bank energies, for each sub-band k and channel m. This algorithm uses filter-bank outputs extracted by the speech recognition system.

- **Random** is a random selection of a channel performed at each utterance.

3.3. Speech recognition

Each of the signals captured by the microphones is decoded with a recognizer implemented with the Kaldi speech recognition toolkit [27], with the following configuration. The language and lexicon models are built according to the s5 recipe included in the Kaldi WSJ configuration. The recognition uses deep neural networks, trained according to Karel's recipe [28], on top of MFCC-LDA-MLLT-fMLLR transformed features. The network architecture is shaped by 6 hidden layers of 1024 neurons, with a context window of 11 consecutive frames (5 before and 5 after the analysis frame), and an initial learning rate of 0.008. The recognition performance on the close-talk material yields a word error rate (WER) of 3.7%.

3.4. Evaluation Methodology

CS approaches are traditionally evaluated by means of recognition results using models trained on clean speech [8, 13] and expressed in terms of WER. Such an experimental setup however introduces certain limitations. First, in a complex task as the WSJ, recognition of distant reverberant speech using clean acoustic models results in a significant performance loss [6]. Evaluating a CS method in such model mismatching cases, makes it hard to individuate any of its possible advantages. Second, commonly exploited approaches that aim at the improvement of the single distant microphone (SDM) recognition accuracy, often result in more distorted channels having a WER as low as, or in certain cases lower than, the least distorted available channel.

In order to obtain a CS evaluation methodology, unrestrained from the above limitations, we propose the use of two evaluation measures, in addition to the WER: (i) the matching rate to an informed CS, and (ii) the average normalized CD between the selected channel and its close-talk reference.

The Informed CS Matching (ICSM) rate is a meaningful evaluation measure in studying how often a certain method succeeds in selecting the least distorted channel. It is defined as

$$ICSM = \frac{\text{\# of matching selections}}{N} \quad , \tag{8}$$

where N is the total number of utterances in the test set. In principle, any objective measure can be used to create the matching ground truth; in this study, we use the informed CS based on CD. It is worth noting that such a matching measure can not be computed using recognition performance because, as previously discussed, more than one channel may achieve the same minimum WER.

The Average Normalized CD (ANCD) is computed as follows:

$$ANCD = \frac{1}{N} \sum_{n} \tilde{d}_n(\boldsymbol{c}_x, \boldsymbol{c}_{\hat{M}}) \quad , \tag{9}$$

where $\boldsymbol{c}_x$ and $\boldsymbol{c}_{\hat{M}}$ are the cepstra vectors of the close-talk signal, x, and the channel, $\hat{M}$, selected by a certain CS method respectively. Here, $\tilde{d}_n(\boldsymbol{c}_x, \boldsymbol{c}_{\hat{M}})$ refers to the cepstral distance, normalized over the maximum CD from all signals, for the utterance n. Therefore, this measure will be bounded within the ANCD of the informed CS method and 1:

$$ANCD \in [\min_m(\tilde{d}_n(\boldsymbol{c}_x, \boldsymbol{c}_m)), 1] \quad . \tag{10}$$

4. Results and Discussion

In this section, we analyze the performance of the proposed CS methods using the previously described corpora and evaluation criteria. First, we present the CD of the 6 different microphones to the close-talk signals. Second, the proposed ICSM rate and ANCD measures are displayed. Third, recognition results are reported.

4.1. Single distant microphone Cepstral Distance

Table 1 reports the average CD between the close-talk signal and each of the SDM used in the study. For the direct simulated case (DirSim), the channel that is intuitively identified as optimal (L2R) has the lowest CD to the close-talk. In the remaining cases, the same trend is not evident because of the

Table 1: Average CD of the distributed microphones.

	Direct		Mixed	
SDM	DirSim	DirReal	MixSim	MixReal
L1C	3.92	2.98	3.79	3.09
L2R	3.25	3.14	3.71	3.15
L3L	3.74	3.05	3.75	3.13
L4L	3.93	3.05	3.81	3.12
LA6	3.78	2.97	3.73	3.04
LD07	3.87	2.89	3.73	3.01

Table 2: Informed CS Matching Rate (ICSM) (%).

CS	DirSim	DirReal	MixSim	MixReal
EV	47.92	31.70	39.36	39.85
CDref	75.00	81.70	75.30	52.32

Table 3: Average Normalized CD (ANCD) between the selected channel and its clean reference.

CS	DirSim	DirReal	MixSim	MixReal
CDi	0.82	0.84	0.85	0.88
EV	0.91	0.89	0.91	0.91
CDref	0.88	0.86	0.89	0.89

averaging among the multiple POSORIs adopted by the speakers. However, in a per utterance analysis, it is clear that when a direct path between the speaker and one of the microphones exists, the corresponding signal has the lowest CD among all the microphones.

4.2. Proposed evaluation

In Table 2, the informed CS matching rate, ICSM, is presented for EV and CDref. The proposed blind CS significantly outperforms both EV and Random CS, which in this experimental setup, for the 6 microphones used, would achieve an ICSM rate of $1/6 \approx 16\%$. CDref achieves a relatively low ICSM rate for the MixReal case, which can be attributed to the fact that this case considers more complex situations, comprising multiple non-direct POSORIs. This type of setup comes in contrast to the original assumption of the proposed method concerning the availability of a direct channel. Moreover, even for an informed CS method such schemes can not be properly addressed, since a selection among highly distorted channels is not always relevant.

In Table 3, the Average Normalized CD, ANCD, is presented for CDi, EV and CDref. It is recalled here that the ANCD of CDi is the upper-bound for a blind CS method. Furthermore, it can be viewed as an indication of the complexity of the conditions of each dataset. As an example, the higher ANCD for the CDi in MixReal evidences the inclusion of more unfavorable cases than in DirReal. This confirms the previously discussed observations concerning the complexity of the MixReal dataset. The proposed blind CS method achieves an average distance closer to the one reached by the informed method, since, as indicated in the ICSM rate evaluation, these two methods repeatedly select the same channel.

4.3. Recognition results

With regard to the recognition performance, Table 4 reports the WER for the recognition of the SDM for each experiment. An interesting observation concerns the low WER achieved by the intuitively best channel (L2R) in the DirSim case. However, there is no direct agreement in the channel ranking given by the objective SDM scoring and the SDM WER. Finally, Table 5 presents the average recognition performance of the CS methods for each dataset. It is recalled here, that Random CS roughly corresponds to the average of the SDM WER. The average CS WER is improved over EV with the proposed blind method for all cases, as shown in Table 5, where the corresponding relative improvement (Rel. Imp.) is reported.

Table 4: WER [%] of the distributed microphones.

SDM	DirSim	DirReal	MixSim	MixReal
L1C	16.6	14.4	16.0	14.8
L2R	10.8	19.2	15.8	16.2
L3L	13.6	15.8	16.5	15.2
L4L	15.0	16.3	17.0	15.1
LA6	16.5	15.1	17.7	14.9
LD07	14.8	14.2	16.4	14.7
Avg	14.5	15.8	16.6	15.2

Table 5: WER [%] by various CS methods.

CS	DirSim	DirReal	MixSim	MixReal
CDi	10.8	12.0	12.8	12.6
EV	12.7	14.7	14.6	13.9
CDref	12.1	12.5	14.1	13.7
Random	14.5	15.9	16.8	15.3
Rel. Imp.	4%	14%	3%	1%

When observing the proposed evaluation measures in addition to the recognition accuracy, one can gain a deeper understanding of the strength of the proposed blind method above the EV based one. See for example the case of DirReal, where for both CS methods WER is reduced in comparison to SDM. However, ICSM rate of CDref is significantly closer to a perfect matching rate, a fact not evidenced from the WER. These remarks indicate the previously discussed gap in the way CS is traditionally evaluated, by means of WER, and the need for evaluation measures similar to the ones introduced in this paper.

5. Conclusions

In this work, we presented a CS framework exploiting the CD, both as a channel scoring function and as a means of detailed evaluation. Through a series of experimental cases, we have proved that the proposed blind CS method (i) improves in all cases the average SDM WER, (ii) consistently outperforms the state-of-the-art EV-based CS method and (iii) successfully selects the least distorted channel when sufficient room coverage is provided by the microphone network. Furthermore, it is illustrated how the standard evaluation of CS, based solely on WER, hides the strengths and weaknesses of different methods. So far, we have considered reverberation to be the main source of degradation of distant speech, however, in a real scenario, environmental noise significantly affects the captured signals. In a future work, different types of noise, and with different SNR, will be studied. In addition, it is interesting to study the use of other objective speech processing measures for CS, both in an informed and blind fashion. Another open topic derived from this study, concerns finding more effective solutions when facing complex conditions, that involve unfavorable speaker positions and/or orientations. A possible direction towards this goal is to detect these cases and replace the CS, given by existing blind methods, with novel techniques.

6. References

[1] M. Wölfel and J. McDonough, *Distant Speech Recognition*. Wiley, 2009.

[2] K. Kinoshita, T. Nakatani, and M. Miyoshi, "Efficient blind dereverberation framework for automatic speech recognition," in *9th European Conference on Speech Communication and Technology*, 2005, pp. 3145–3148.

[3] K. Kumatani, T. Gehrig, U. Mayer, E. Stoimenov, J. McDonough, and M. Wolfel, "Adaptive beamforming with a minimum mutual information criterion," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 15, no. 8, pp. 2527–2541, 2007.

[4] P. A. Naylor and N. D. Gaubitch, "Speech dereverberation," in *Tutorial presented at European Signal Processing Conference*, 2010.

[5] S. Makino, T. Lee, and H. Sawada, *Blind Speech Separation*, ser. Signals and Communication Technology. Springer, 2010. [Online]. Available: http://books.google.com/books?id=2IuxcQAACAAJ

[6] J. Barker, R. Marxer, E. Vincent, and S. Watanabe, "The Third CHiME Speech Separation and Recognition Challenge: Dataset, task and baselines," in *2015 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU 2015)*, 2015.

[7] M. Brandstein and D. Ward, *Microphone Arrays: Signal Processing Techniques and Applications*. Springer, 2001.

[8] M. Wolf and C. Nadeu, "Channel selection measures for multimicrophone speech recognition," *Speech Communication*, vol. 57, pp. 170–180, 2014.

[9] ——, "Towards microphone selection based on room impulse response energy-related measures," in *Proc. of I Joint SIG-IL/Microsoft Workshop on Speech and Language Technologies for Iberian Languages*, Porto Salvo, Portugal, 2009, pp. 61–64.

[10] M. Wölfel, C. Fügen, S. Ikbal, and J. W. McDonough, "Multisource far-distance microphone selection and combination for automatic transcription of lectures." in *INTERSPEECH*. Citeseer, 2006.

[11] M. Wolf and C. Nadeu, "On the potential of channel selection for recognition of reverberated speech with multiple microphones," in *INTERSPEECH*, Tokyo, Japan, 2010, pp. 80–83.

[12] K. Kumatani, J. McDonough, J. F. Lehman, and B. Raj, "Channel selection based on multichannel cross-correlation coefficients for distant speech recognition," in *Hands-free Speech Communication and Microphone Arrays, 2011 Joint Workshop on*. IEEE, 2011, pp. 1–6.

[13] I. Himawan, P. Motlicek, S. Sridharan, D. Dean, and D. Tjondronegoro, "Channel selection in the short-time modulation domain for distant speech recognition," in *Proceedings of Interspeech*, no. EPFL-CONF-209075, 2015.

[14] M. Wolf, "Channel selection and reverberation-robust automatic speech recognition," *PhD, Universitat Politècnica de Catalunya*, 2013.

[15] A. Gray and J. Markel, "Distance measures for speech processing," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 24, no. 5, pp. 380–391, Oct 1976.

[16] N. Kitawaki, H. Nagabuchi, and K. Itoh, "Objective quality evaluation for low-bit-rate speech coding systems," *IEEE Journal on Selected Areas in Communications*, vol. 6, no. 2, pp. 242–248, Feb 1988.

[17] S. Furui and M. M. Sondhi, *Advances in Speech Signal Processing*, ser. Electrical and Computer Engineering. Marcel Dekker Inc., 1991.

[18] K. Kinoshita, M. Delcroix, T. Yoshioka, T. Nakatani, A. Sehr, W. Kellermann, and R. Maas, "The REVERB challenge: A common evaluation framework for dereverberation and recognition of reverberant speech," in *Applications of Signal Processing to Audio and Acoustics, 2013 IEEE Workshop on*. IEEE, 2013, pp. 1–4.

[19] Y. Hu and P. C. Loizou, "Evaluation of objective quality measures for speech enhancement," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 16, no. 1, pp. 229–238, 2008.

[20] L. R. Rabiner and R. W. Schafer, *Theory and application of Digital Speech Processing*. PEARSON, 2011.

[21] X. Huang, A. Acero, and H.-W. Hon, *Spoken language processing: A guide to theory, algorithm, and system development*. Prentice Hall PTR, 2001.

[22] L. Rabiner and B.-H. Juang, "Fundamentals of speech recognition," 1993.

[23] H. Kuttruff, *Acoustics: An Introduction*. CRC Press, 2007.

[24] L. Cristoforetti, M. Ravanelli, M. Omologo, A. Sosi, A. Abad, M. Hagmüller, and P. Maragos, "The DIRHA simulated corpus." in *9th edition of the Language Resources and Evaluation Conference*, Reykjavik, Iceland, 2014, pp. 2629–2634.

[25] J. Garofalo, D. Graff, D. Paul, and D. Pallett, "Continous speech recognition (CSR-I) Wall Street Journal (WSJ0) News Complete," *LDC93S6A. DVD. Linguistic Data Consortium, Philadelphia*, 1993.

[26] M. Ravanelli, L. Cristoforetti, R. Gretter, M. Pellin, A. Sosi, and M. Omologo, "The DIRHA-English corpus and related tasks for distant-speech recognition in domestic environments," in *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*. IEEE, 2015, pp. 275–282.

[27] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz *et al.*, "The Kaldi speech recognition toolkit," in *IEEE 2011 workshop on automatic speech recognition and understanding*, no. EPFL-CONF-192584. IEEE Signal Processing Society, 2011.

[28] K. Veselỳ, A. Ghoshal, L. Burget, and D. Povey, "Sequence-discriminative training of deep neural networks." in *INTERSPEECH*, 2013, pp. 2345–2349.