Channel Selection in the Short-time Modulation Domain for Distant Speech Recognition

Ivan Himawan[1]*, Petr Motlicek*[1]*, Sridha Sridharan*[2]*, David Dean*[2]*, Dian Tjondronegoro*[2]

[1]Idiap Research Institute, Martigny, Switzerland
{ihimawan,motlicek}@idiap.ch
[2]Science and Engineering Faculty, Queensland University of Technology
{s.sridharan,d.dean,dian}@qut.edu.au

Abstract

Automatic speech recognition from multiple distant microphones poses significant challenges because of noise and reverberations. The quality of speech acquisition may vary between microphones because of movements of speakers and channel distortions. This paper proposes a channel selection approach for selecting reliable channels based on selection criterion operating in the short-term modulation spectrum domain. The proposed approach quantifies the relative strength of speech from each microphone and speech obtained from beamforming modulations. The new technique is compared experimentally in the real reverb conditions in terms of perceptual evaluation of speech quality (PESQ) measures and word error rate (WER). Overall improvement in recognition rate is observed using delay-sum and superdirective beamformers compared to the case when the channel is selected randomly using circular microphone arrays.

Index Terms: channel selection, signal quality, microphone arrays, reverberation

1. Introduction

Close talking microphones give the best signal quality and produce the highest accuracy from the current automatic speech recognition (ASR) systems but their use is obtrusive. Employment of microphone arrays in contrast to close talking microphones alleviates the feeling of discomfort and distraction to the user. For this reason, microphone arrays are popular and have been used in a wide range of applications such as teleconferencing, hearing aids, speaker tracking, and as the front-end to ASR systems. However, their performance tends to decrease as the distance from microphones to the speaker's mouth increases in which noise and reverberation dominates the direct sound [1].

In the case of multi-microphone approaches, selecting a subset of microphones for beamforming could dramatically improve the performance of speech enhancement and ASR systems. This is particularly useful when microphones are spatially distributed in user's environment. The subset of microphones could be selected on the basis of a stronger peak in the cross correlation function by assuming that signals of reliable channels are often correlated with each others [2]. Also, different measures such as intra-clusters distances and their promixities to a speaker could be used to form clusters of microphones [3]. Another approach to selecting microphones which does not require a spatial structure of the microphone set is by employing channel selection measures. The channels which are deemed to have sufficient quality can be selected for further processing such as beamforming or as an input to ASR systems.

In general, the measures for the channel selection approaches can be categorized into two groups. The first is the signal-based measures. The signal-based measures use signal processing techniques to identify the least distorting channel and operates in the front-end of the ASR system. As the acoustic wave propagates from the sound source, its amplitude is decaying at a rate proportional to the distance from the source. Hence, the sound energy received by the closest microphone is presumably stronger compared to microphones that are located further away. This leads to a straightforward way to identify the least distorting channel by calculating the signal energy relative to other microphones and has been reported to achieve good results [4]. The issue with this method is that the perfect calibration may be necessary for all microphones because of a variation in microphone responses (i.e. gain and frequency responses). Another measure such as signal-to-noise ratio (SNR) may also be used. This requires voice activity detection to estimate noise power [5]. The SNR may not be a reliable indicator signal quality for speech signal recorded by distant-talking microphones where reverberation dominates the energy of the original signal [6]. The second measures for the channel selection approach are the decoder-based measures [7, 5]. These measures involve some kind of classification in the decoding part of recognition system such as selecting channel with the maximum acoustic likelihood [7]. One of the drawback of the decoder based measures is that the recognition must first take place before any channel can be selected which make these measures to be more computationally demanding.

The speech degraded by reverberation is usually modeled by the convolution of the room impulse response (RIR) with the original speech signal. Hence, the correlation values between different RIR features and the word error rate can be used to predict recognition performance before the speech recognition takes place. Assuming an exact knowledge of RIR, such measure can then be used for selecting the best microphone before entering the recognition system [4]. Unfortunately, the RIR estimate is not always available and the distortion must be measured from the recorded speech signal directly. Because reverberation results in the temporal smearing of the short-time spectra, [6] used the estimates of the variance of compressed filter bank energies to select channels which give the highest energy for all sub-bands as the least distorted channel.

Previous research has shown that the modulation frequen-

This work was supported by the European Community under the project DBox: A generic dialog box for multi-lingual conversational applications. This work was also partially supported by the EC FP7 funding, under "Speaker Identification Integrated Project (SIIP)".

cies that is in the range between 4 and 16 Hz contribute the most to intelligibility, with spectral peaks at approximately 4 Hz, corresponding to the rate of syllables from the spoken speech [8, 9]. Because the background noise reduces the depth of low-frequency envelope modulations [10, 11] and reverberation to induce a multiplicative distortion in the modulation spectral domain [12], these facts can be used to predict whether the recorded speech has been influenced by noise and reverberation. The measure proposed in this paper based on the assumption that clean speech has more modulation than noisy or reverberated speech which is formulated as the ratio of energy between the microphone channel and beamformed output in the short-term modulation spectrum domain. Similar task but different approach in modulation spectrum has been attempted by selecting a channel in which the normalized modulation energy of the area between 0.25Hz to 16Hz is maximum [4]. The proposed technique is analogous to signal-to-reverberant (SRR) criterion for sub-band channel selection in the acoustic-frequency domain [13]. Instead of using clean signal as a reference and a reverberant signal as a target signal in the SRR computation, the proposed method assigns signal in each microphone channel as a reference and a beamformed signal will serve as the target signal in the SRR computation.

The frame based compensation techniques for ASR such as cepstral mean normalization are motivated by assumption that linear channel distortion (e.g., due to reverberation) which is convolutive in the time domain can be considered as additive noise in the log-spectral domain [14]. Although the feature processing pipeline of ASR system has attempted to normalize the effect of reverberation, the proposed channel selection can be used to select reliable channels for beamforming. This will be useful if speakers move their positions and in ad-hoc array situations.

Experiments in this paper are conducted using the single speaker portions of the Multi-Channel Wall Street Journal Audio Visual (MC-WSJ-AV) corpus [15], which offers an intermediate task between simple digit recognition and large vocabulary conversational speech recognition. The corpus' recordings which are recorded at the University of Edinburgh are made using two small circular arrays for six conditions in which the speaker reads sentences from six different positions within the meeting room. The reverberation time of this room is approximately 0.7s [16]. The remainder of this paper is organized as follows. Section 2 describes the framework for signal processing in the short-term modulation domain followed by the proposed modulation spectrum based channel selection method. Sections 3 and 4 present and discuss experiments on the MC-WSJ-AV corpus, followed by conclusions in Section 5.

2. Modulation Domain Processing

The proposed channel selection method uses a dual analysis-modification-synthesis framework which allow access to the short-time modulation spectral domain [17, 11]. Note that for our case, only signal analysis is performed without signal modification and reconstruction. Under this framework, the speech signal is processed framewise using short-time Fourier analysis and the time trajectories of the acoustic magnitude spectrum (accumulated over a finite interval of Ts at fixed acoustic frequencies) are subjected to a second short-time Fourier analysis to produce the modulation spectrum.

For a discrete-time signal $x(n)$, the short-time Fourier transform (STFT) is given by:

$$X(n, f) = \sum_{l=-\infty}^{\infty} x(l)w(n-l)\exp^{-j2\pi fl/N}, \quad (1)$$

where n refers to the discrete-time index, f is the index of the discrete acoustic frequency, N is the acoustic frame duration (in samples), and $w(n)$ is the acoustic analysis window function. Here, a Hamming window is used as the analysis window function. In polar form, the STFT of the speech signal can be written as:

$$X(n, f) = |X(n, f)| \exp^{j\angle X(n,f)}, \quad (2)$$

where $|X(n, f)|$ denotes the acoustic magnitude spectrum and $\angle X(n, f)$ denotes the acoustic phase spectrum.

The modulation spectrum for a given frequency is calculated as the STFT of the time series of the acoustic spectral magnitudes at that frequency. Hence, the modulation spectrum is calculated as follows:

$$\chi(\eta, f, m) = \sum_{l=-\infty}^{\infty} |X(l, f)|\nu(\eta - l)\exp^{-j2\pi ml/M}, \quad (3)$$

where η is the acoustic frame number, f refers to the index of the discrete-acoustic frequency, m refers to the index of the discrete modulation frequency, M is the modulation frame duration, and $\nu(\eta)$ is the modulation analysis window function.

In polar form, the modulation spectra can be written as:

$$\chi(\eta, f, m) = |\chi(\eta, f, m)| \exp^{j\angle\chi(\eta,f,m)}, \quad (4)$$

where $|\chi(\eta, f, m)|$ is the modulation magnitude spectrum, and $\angle\chi(\eta, f, m)$ is the modulation phase spectrum. In the following the dependencies on η is omitted for lucidity.

2.1. Modulation Spectrum based Channel Selection

The proposed measure is formulated as the ratio of instantaneous measurements between the signal from each microphone and the beamformed output in the short-time modulation spectrum domain, defined as:

$$\zeta_c(f, m) = 10log_{10}\frac{|\chi_c(f, m)|^2}{|B(f, m)|^2}, 0 \le m \le M, \quad (5)$$

where $\chi_c(f, m)$ and $B(f, m)$ denote the modulation spectra of microphone channel c and beamforming signal respectively, and M denotes the highest modulation frequency. The $B(f, m)$ is obtained using the signal processing steps to obtain modulation spectrum (i.e., instead of $x(n)$ in Equation 1, the delay-sum beamforming output is used).

The microphone channels with $\zeta_c(f, m)$ greater than threshold θ are selected as the best channels. In this paper, this information is aggregated across frequency and modulation bins, and across frames for every available channels, and channels which give the highest scores are selected as the best channels. It is possible to set the range of modulation frequencies with cutoff frequencies of M_c in Equation 5 ($M_c = M$) over which the channel selection is to be performed.

3. Experiments

3.1. Database Specifications

Experiments were conducted on a subset of MC-WSJ-AV corpus. Only the single-speaker stationary sentences were used.

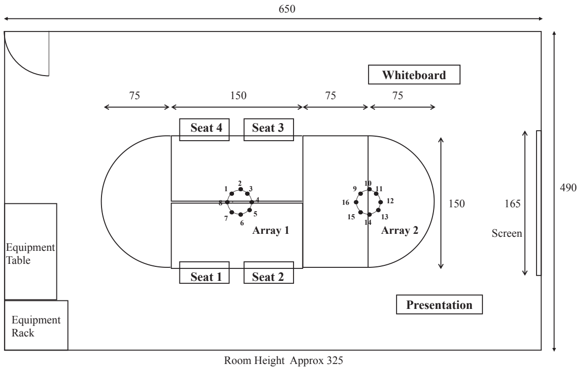

Figure 1: *The layout of the Edinburgh Meeting Room according to [15]. The four reading positions are indicated as Seat 1, Seat 2, Seat 3, and Seat 4.*

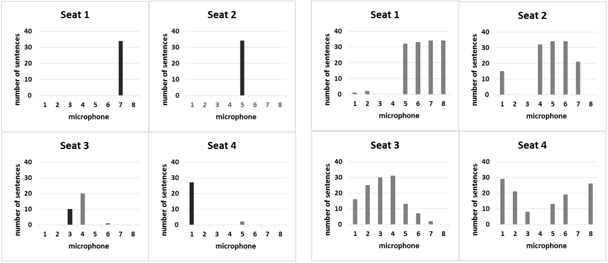

Figure 2: *The best (left) and four best channels (right) obtained from the modulation channel selection for the four speaking positions from circular array 1. The darker bar indicates manually chosen closest microphone to the speaker based on Figure 1.*

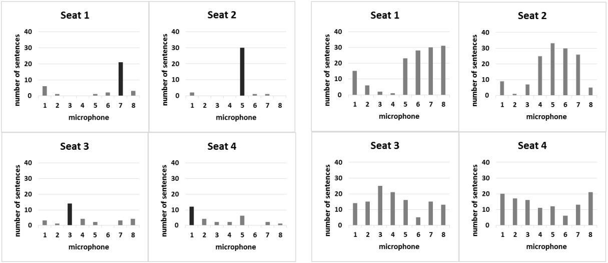

Figure 3: *The best (left) and four best channels (right) obtained from the energy-based channel selection for the four speaking positions from circular array 1.*

In the single-speaker stationary task, there are six conditions in which the speaker reads sentences from six different positions within the meeting room. Only four seating conditions with a total of 128 utterances were used for experiments in this paper: speaker sits at seat 1 (Seat 1) with the total number of 34 sentences, speaker sits at seat 2 (Seat 2) with the total number of 34 sentences, speaker sits at seat 3 (Seat 3) with the total number of 31 sentences, and speaker sits at seat 4 (Seat 4) with the total number of 29 sentences. Two array geometries on which the proposed method is tested: (1) circular array 1 - a fixed 8-element, equally spaced, circular microphone array with a diameter of 20cm (denoted as Array 1 using microphones 1 to 8 in Figure 1), and (2) circular array 2 - with a similar geometry and an equal number of elements to array 1 (denoted as Array 2 using microphones 9 to 16 in Figure 1).

3.2. Channel Selection Experiments

The modulation spectrum based channel selection stimuli were constructed with an acoustic frame duration set to 32 ms and the modulation frame duration set to 256 ms. A 75% overlap was used between frames. The modulation threshold θ set to -5dB with the modulation cutoff frequency M_c set to 16Hz. For each array geometry on which the channel selection algorithms was tested, the beamforming modulation spectrum $|B(f,m)|$ in Equation 5 is computed from beamformed output of that array. The best channel is selected from that array of microphones which give the highest $\zeta_c(f,m)$ value. In similar fashion, the four best microphones can be selected by finding four microphone channels with the highest ζ_c. The proposed method is compared with the energy-based measure (which select microphones with the highest energy relative to others). Microphones within an array are calibrated to have similar gain level before being processed by the proposed and energy-based methods.

Table 1: PESQ measures averaged for every microphone (mic. 1-16) for each speaking position. The figures in bold show the best channel using the proposed method with the highest number of sentences selected for circular array 1 (mic. 1-8) and 2 (mic. 9-16).

Mic.	Seat 1	Seat 2	Seat 3	Seat 4
1	2.19	2.13	2.08	**2.18**
2	2.16	2.13	2.11	2.16
3	2.14	2.14	2.15	2.12
4	2.14	2.16	**2.17**	2.11
5	2.16	**2.16**	2.16	2.12
6	2.17	2.13	2.14	2.13
7	**2.18**	2.11	2.13	2.14
8	2.20	2.12	2.11	2.16
9	1.99	2.11	**2.13**	**1.98**
10	1.96	2.07	2.10	1.97
11	1.93	2.06	2.06	1.95
12	1.92	2.06	2.04	1.96
13	1.94	2.10	2.05	1.98
14	1.98	2.14	2.06	2.01
15	**2.00**	**2.16**	2.10	2.02
16	2.01	2.15	2.12	2.00

Table 2: PESQ measures from modulation spectrum based channel selection (MODS) using circular array 1 and 2 in four speaking positions. As a comparison, the PESQ measures from the best channel using energy-based measure (ENER) are presented. The results are averaged over all utterances for each speaking position.

	Array 1		Array 2	
Spk.	MODS	ENER	MODS	ENER
Seat 1	2.18	2.20	1.99	1.97
Seat 2	2.16	2.15	2.16	2.13
Seat 3	2.16	2.14	2.12	2.12
Seat 4	2.18	2.17	2.00	2.01

Table 3: WERs[%] on the evaluation set of MC-WSJ-AV corpus: RND refers to randomly selected microphone. MODS refers to the proposed technique. ENER refers to the energy-based method.

	Array 1			Array 2		
Spk.	RND	MODS	ENER	RND	MODS	ENER
Seat 1	47.1	44.7	47.6	74.7	73.9	74.1
Seat 2	44.7	38.9	39.5	62.7	54.9	59.9
Seat 3	44.2	40.1	44.0	58.4	53.5	53.5
Seat 4	43.1	36.4	36.2	68.4	59.0	66.1

Table 4: WERs[%] on the evaluation set of MC-WSJ-AV corpus: RND DS and RND SD refer to delay-sum and superdirective beamforming using 4 randomly selected microphones respectively. MODS DS and MODS SD refer to delay-sum and superdirective beamforming using 4 selected microphones from the proposed technique. The last column of the table shows the performance of delay-sum beamforming using all microphones from both arrays.

	Array 1				Array 2				Both Arrays
Spk.	RND DS	MODS DS	RND SD	MODS SD	RND DS	MODS DS	RND SD	MODS SD	DS
Seat 1	35.5	34.1	27.7	26.7	70.5	70.5	76.5	70.6	44.0
Seat 2	27.7	25.7	22.2	21.0	44.4	43.6	42.6	37.4	22.3
Seat 3	25.3	27.1	22.4	20.5	48.9	43.4	47.9	44.0	24.2
Seat 4	25.3	22.9	19.9	18.7	60.4	55.6	71.0	63.2	33.4

The proposed method is evaluated for each sentence recording for the four speaking positions. The selected best channel for each sentence is accumulated and shown as bar plots on the left side of Figure 2 for circular array 1. Since the best channel selected by the algorithm can be different for each utterance, more than one best channel can be selected for the best channel in which the number of sentences for each selected channel correspond to the height of the bar. In similar way, the best four channels for each utterance are accumulated and shown as bar plots on the right side of the same figure. For the energy-based method, the results are shown as bar plots in Figure 3 for the best and four best channels for circular array 1.

The perceptual objective measure ITU-T Rec. P.862 PESQ is also used for evaluating the speech quality of selected channels. The PESQ is an intrusive-based method which predicts the speech quality using the clean speech signal as a reference and compare it with the distorted signal. In this paper, the PESQ score for each microphone is measured using the headset microphone signal as a reference and the output is expressed in terms of mean opinion score (MOS) with high values indicating better quality. For experiments in this paper, the PESQ software [18] was used to predict the mean opinion score. Table 1 shows PESQ scores for every microphone (mic. 1-16) and for each speaking position. The PESQ scores for the proposed and the energy-based method are presented in Table 2 for circular array 1 and 2.

The speech recognition experiments are conducted when the proposed approach is used as a front-end for ASR systems. In this paper, the ASR system employs hybrid HMM/DNN acoustic model trained from 18.9 hours clean speech data from WSJCAM0 using KALDI speech recognition toolkit [19, 20]. The baseline performance on the headset recording of the MC-WSJ-AV with a total of 128 utterances yields a WER of 6.1% with a highly-pruned trigram language model. All speech recognition results quoted in this paper are the percentage of word error rate (WER). Table 3 shows the WERs of the best channel from the proposed and energy-based measures and if the channel is selected randomly. The results using delay-sum and superdirective beamformers of the best four microphones are presented in Table 4.

4. Discussion

Using a circular array 1 with 8-elements, Figure 2 shows that the proposed algorithm selects mostly the spatially closest microphone to the speaker with a higher accuracy for all seating positions. The best microphone for Seat 1 and 2 are microphone number 7 and 5 respectively. Similarly, the best microphone for Seat 4 is the closest microphone 1 with a few number of instances where microphone 5 is selected. For Seat 3, two spatially closest microphone to the speaker are chosen which are microphone 4 (the highest) and microphone 3. Note that microphone 3 and 4 are located spatially next to each other and the actual distance from both microphones to the speaker may roughly similar.

The best channel obtained from the modulation channel selection for circular array 2 are not shown in this paper due to the space limitation. In terms of PESQ as shown in Table 1, the best microphones with the highest number of sentences selected for circular array 2 are generally have higher scores compared to other microphones. Similar trends are also shown for the best channels from circular array 1. Note that very small differences in PESQ scores between microphones are because of the similar quality microphones used are located spatially close.

The simple energy-based channel selection is not as reliable as the proposed method for selecting the best and the four best microphones. Figure 3 shows that compared to the proposed method, more microphones which have lower PESQ scores are considered as the best channels. The results are worse for selecting the best channel for Seat 3 and 4. From results in Table 2, in most seating conditions, the channels selected from the proposed method give better or equal performance compared to energy-based method.

In terms of WER as shown in Table 3, the overall performance obtained by the proposed method is better compared to enery-based and random selections for circular array 1 and 2. Overall, using four microphones for beamforming with the proposed method allow improvements for both circular array 1 and 2 compared to random microphone selection as shown in Table 4. Note that using all 16-microphones for beamforming does not necessarily give the best performance compared to using only 4-microphones for Seat 1 and 4. In ad-hoc array situations where microphones are distributed in user's environment, selecting microphones closest to the speaker will be beneficial.

The worse performance of superdirective beamforming compared to delay-sum beamforming for circular array 2 is due to the error in delay estimation and the high sensitivity of the beamformer with such deviations [21]. In particular, no improvement is shown for Seat 1 and Seat 4 (i.e, the two positions which are located furthest from circular array 2). Nevertheless, the proposed method is better compared to random selection.

5. Conclusions

This paper presents method for selecting reliable channels based on selection criterion operating in the short-time modulation domain. The evaluations on speech captured from distant talking microphones show that the developed criterion capable of selecting microphones of higher speech quality as indicated by PESQ measures and WER for closely-spaced array such as a circular array. Future works include investigating the algorithm proposed here to the situation where speakers are moving and developing an automatic method to determine the optimum number of channels using ad-hoc microphone arrays.

6. References

[1] J. Bitzer, Klaus Uwe Simmer, and Karl-Dirk Kammeyer, "Multi-microphone noise reduction techniques as front-end devices for speech recognition," *Speech Communication*, 2001.

[2] K. Kumatani, J. McDonough, J. Lehman, and B. Raj, "Channel selection based on multichannel cross-correlation coefficients for distant speech recognition," in *Proceedings of Hands-free speech communication and microphone arrays*, May 2011, pp. 1–6.

[3] I. Himawan, I. McCowan, and S. Sridharan, "Clustered blind beamforming from ad-hoc microphone arrays," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 661–676, May 2011.

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

[5] M. Wölfel, C. Fgen, S. Ikbal, and J. W. Mcdonough, "Multi-source far-distance microphone selection and combination for automatic transcription of lectures," in *Proceedings of Interspeech*, 2006.

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

[7] Y. Shimizu, S. Kajita, K. Takeda, and F. Itakura, "Speech recognition based on space diversity using distributed multi-microphone," in *Proceedings of ICASSP*, 2000, pp. 1747–1750.

[8] R. Drullman, J. M. Festen, and R. Plomp, "Effect of reducing slow temporal modulations on speech reception," *Journal of the Acoustical Society of America*, vol. 95, pp. 2670–2680, 1994.

[9] T. Arai, M. Pavel, H. Hermansky., and C. Avendano, "Intelligibility of speech with filtered time trajectories of spectral envelopes," in *Proceedings of ICSLP*, 1996, pp. 2490–2493.

[10] X. Xiao, E. S. Chng, and H. Li, "Normalization of the speech modulation spectra for robust speech recognition," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 16, pp. 1662–1674, 2008.

[11] K. Wojcicki and P. Loizou, "Channel selection in the modulation domain for improved speech intelligibility in noise," *The Journal of the Acoustical Society of America*, vol. 131, pp. 2904–2913, 2012.

[12] Bengt J. Borgstrom and Alan McCree, "The Linear Prediction Inverse Modulation Transfer Function (LP-IMTF) Filter for Spectral Enhancement, with Applications to Speaker Recognition," *in Proceedings of IEEE International Conference on Acoustic, Speech and Signal Processing*, 2012.

[13] O. Hazrati and P. C. Loizou, "Tackling the combined effects of reverberation and masking noise using ideal channel selection," *Journal of Speech, Language, and Hearing Research*, vol. 55, pp. 500–510, 2012.

[14] T. Yoshioka, A. Sehr, M. Delcroix, K. Kinoshita, R. Maas, T. Nakatani, and W. Kellermann, "Making Machines Understand Us in Reverberant Rooms: Robustness Against Reverberation for Automatic Speech Recognition," *IEEE Signal Processing Magazine*, Nov. 2012.

[15] M. Lincoln, I. McCowan, J. Vepa, and H. Maganti, "The multi-channel wall street journal audio visual corpus (MC-WSJ-AV): Specification and initial experiments," *in Proc. ASRU*, pp. 357–362, 2005.

[16] Keisuke Kinoshita et al., "Reverb challenge - Evaluating de-reverberation and ASR techniques in reverberant environments [Online]," Internet: http://reverb2014.dereverberation.com/ [March 26, 2015], 2014.

[17] K. Paliwal, B. Schwerin, and K. Wjcicki, "Role of modulation magnitude and phase spectrum towards speech intelligibility," *Speech Communication*, vol. 53, no. 3, pp. 327 – 339, 2011.

[18] P. C. Loizou, *Speech Enhancement: Theory and Practice*. CRC Press, 2007.

[19] Petr Motlicek and Philip N. Garner and Namhoon Kim and Jeongmi Cho, "Accent Adaptation Using Subspace Gaussian Mixture Models," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal processing*, 2013.

[20] Daniel Povey et al., "The Kaldi speech recognition toolkit," in *Automatic Speech Recognition and Understanding*, 2011.

[21] H. L. V. Trees, *Optimum Array Processing - Part IV of Detection, Estimation, and Modulation Theory*. New York: Wiley, 2002.