IDIAP RESEARCH REPORT

BROADBAND BEAMPATTERN FOR MULTI-CHANNEL SPEECH ACQUISITION AND DISTANT SPEECH RECOGNITION

Mohammad J. Taghizadeh Philip N. Garner
Hervé Bourlard

Idiap-RR-39-2011

DECEMBER 2011

Centre du Parc, Rue Marconi 19, P.O. Box 592, CH - 1920 Martigny
T +41 27 721 77 11 F +41 27 721 77 12 info@idiap.ch www.idiap.ch

BROADBAND BEAMPATTERN FOR MULTI-CHANNEL SPEECH ACQUISITION AND DISTANT SPEECH RECOGNITION

Mohammad J. Taghizadeh[1,2], *Philip N. Garner*[1] *and Hervé Bourlard*[1,2]

[1]Idiap Research Institute, Martigny, Switzerland
[2]Ecole Polytechnique Fédérale de Lausanne (EPFL), Switzerland
{mtaghizadeh,pgarner,hbourlard}@idiap.ch

ABSTRACT

Spatial filtering is the fundamental characteristic of microphone array based signal acquisition which plays an important role in applications such as speech enhancement and distant speech recognition. In the array processing literature, this property is formulated upon beam-pattern steering and it is characterized for narrowband signals.

This paper proposes to characterize the microphone array broadband beam-pattern based on the average output of a steered beamformer for a broadband spectrum. Relying on this characterization, we derive the directivity beam-pattern of delay-and-sum and superdirective beamformers for a linear as well as a circular microphone array. We further investigate how the broadband beam-pattern is linked to speech recognition feature extraction; hence, it can be used to evaluate distant speech recognition performance. The proposed theory is demonstrated with experiments on real data recordings.

Index Terms: Broadband beam-pattern, directivity, Delay-and-sum beamformer, Superdirective beamformer, Distant speech recognition

1. INTRODUCTION

Microphone array based speech acquisition is a common practice in hands-free interfaces in state-of-the-art applications such as distant speech recognition [1], speech separation and speaker localization [2, 3]. Multi-channel signal acquisition relies on beamforming or spatial filtering for directional discrimination, and space-time filtering of the signals in the acoustic scene [4, 5]. An important issue then is to design an optimal microphone array to achieve a desired look-angle directivity and suppression of interference. This task is usually entangled with constraints on Signal-to-Noise Ratio (SNR), side-lobe level and beamwidth. Whilst these parameters have been characterized in the array processing literature, the theory usually revolves around narrowband assumptions. In previous art to address the issues in acquisition of wideband signals, Coleman et al. [6] propose to characterize the wideband beam-pattern based on random-process autocorrelations and cross correlations. Their formulation provides mean-square-error and average-gain measures for far-field beamforming obtained through convex optimization. In the microphone array signal processing literature, the broadband characterization of beampattern has not been addressed and, to the extent of our knowledge, the microphone array acquisition has been understood through the narrowband properties [7].

In this paper we formulate the broadband beam-pattern and directivity for acquisition of speech signals. Our formulation exploits the concept of the narrowband beam-pattern while the beamformer's output power for a broadband spectrum is used to characterize the power-pattern. We then show how this formulation is linked to the speech recognition front-end processing, hence the power-pattern enables us to evaluate the performance of the distant speech recognition system.

The rest of the paper is organized as follows: we give a quick overview of the narrowband formulation of the beam-pattern in Section 2.1. The extension of this formulation for speech signal is explained in Section 2.2. Section 2.3 is dedicated to the simulations of the proposed method. Section 3.1 present the experimental demonstration of the theory. The link between the broadband beam-pattern and the speech recognition front-end processing is shown in Sections 3.2 and 3.3. The conclusions are drawn in Section 4.

2. BROADBAND BEAMPATTERN

2.1. Microphone Array Pattern

We consider a general array of isotropic elements in a homogenous medium. A plane wave of the external signal field impinges on the array from the direction of $\bar{a}(\theta,\phi)$ where θ and ϕ denote the elevation and the azimuth angles in spherical coordinates. Wavenumber k is defined as

$$k = \frac{2\pi}{\lambda}\bar{a}(\theta,\phi), \tag{1}$$

where λ is the wavelength in radians with frequency ω. The narrowband beam-pattern is defined as the spatial and temporal frequency response of the array evaluated against the direction [8, 9] and stated as

$$B(\omega,\bar{a}(\theta,\phi)) = \sum_{m\in M} H_m(\omega)e^{-jk.m} \tag{2}$$

where $H_m(\omega)$ is frequency response filter of the microphone located at m; M is the set of microphone locations. The power-pattern is then defined as

$$P(\omega,\bar{a}(\theta,\phi)) = |B(\omega,\bar{a}(\theta,\phi))|^2. \tag{3}$$

The array directivity denoted by D is defined as the ratio of maximum power-pattern to the average of power-pattern in all directions, stated concisely as

$$D(\omega,\bar{a}(\theta_0,\phi_0)) = \frac{P(\omega,\bar{a}(\theta_0,\phi_0))}{\frac{1}{4\pi}\int_{-\pi/2}^{\pi/2} d\theta \int_0^{2\pi} d\phi \sin(\theta) P(\omega,\bar{a}(\theta,\phi))}, \tag{4}$$

where $\bar{a}(\theta_0,\phi_0)$ is the steering direction, which is constant along frequency bands.

2.2. Broadband Beampattern for Speech Acquisition

In this section, we exploit the concept of narrowband beam-pattern and formulate the beam-pattern for acquisition of broadband signals such as speech. Suppose that $S(\omega)$ is the spectral representation of the clean speech signal in Fourier domain (estimated from a database). The spectrum of speech has a complex structure. Most of the energy is generated during the voiced parts and concentrated in three to four formants up to 2 KHz in frequency. So we considered this structure for extracting the beam-pattern. The spectrum of the speech signal can be extracted by the Welch's method. Hence, the response of the array or the beamformer (e.g., superdirective, delay-and-sum) denoted by $F(\omega, \bar{a}(\theta, \phi))$ to the plane wave $S(\omega)$ would be

$$Y(\omega, \bar{a}(\theta, \phi)) = F(\omega, \bar{a}(\theta, \phi))S(\omega). \quad (5)$$

In other words, $Y(\omega)$ is the beamformer output for the broadband spectrum of the signal over the sphere of look directions. Given $Y(\omega)$, we define the broadband beam-pattern as

$$B_{sp}(\bar{a}(\theta, \phi)) = \frac{\sqrt{\int_0^{\omega_N} Y^2(\omega, \bar{a}(\theta, \phi))d\omega}}{\sqrt{\int_0^{\omega_N} Y^2(\omega, \bar{a}(\theta_0, \phi_0))d\omega}}, \quad (6)$$

where ω_N is the Nyquist frequency. The proposed beam-pattern can be interpreted as a weighted average of the beamformer's output over the speech signal. Thereby, it is mostly influenced by the dominant frequencies of speech spectrum. Accordingly, the power-pattern for the broadband spectrum would be

$$P_{sp}(\bar{a}(\theta, \phi)) = |B_{sp}(\bar{a}(\theta, \phi))|^2, \quad (7)$$

and the directivity for speech acquisition will be defined as

$$D_{sp}(\bar{a}(\theta_0, \phi_0)) = \frac{P_{sp}(\bar{a}(\theta_0, \phi_0))}{\frac{1}{4\pi}\int_{-\pi/2}^{\pi/2} d\theta \int_0^{2\pi} d\phi \sin(\theta) P_{sp}(\bar{a}(\theta, \phi))}. \quad (8)$$

The directivity as defined above, can be interpreted as the array gain for speech acquisition in the presence of isotropic noise. The 3dB beamwidth is a measure of the width of the main lobe of the beam-pattern. It is defined as the maximum angle in normalized power-pattern for which power is above 0.5. Applying the normalization in weights so that $P_{sp}(\bar{a}(\theta_0, \phi_0)) = 1$, the normalized directivity can be written as

$$\hat{D}_{sp}(\bar{a}(\theta_0, \phi_0)) = \left(\frac{1}{4\pi} \int_{-\pi/2}^{\pi/2} d\theta \int_0^{2\pi} d\phi \sin(\theta) P_{sp}(\bar{a}(\theta, \phi) \right)^{-1}. \quad (9)$$

The broadband directivity provides an objective for the acquisition of the speech signal. Assuming that the background noise is diffuse, SNR maximization becomes equal to the maximization of the broadband directivity.

2.3. Simulations

We present some empirical studies on the proposed theory. These studies aim to provide a broad view of the beam-pattern for different acquisition set-up of microphone array and to compare and contrast the speech vs. narrowband signal's beam-pattern.

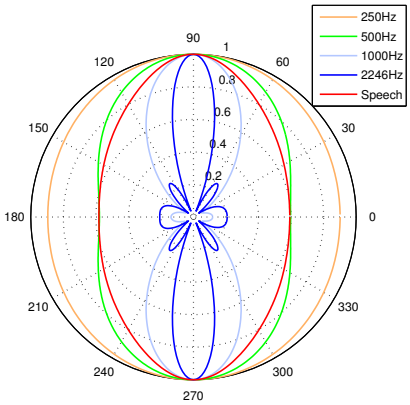

Fig. 1. Speech vs. narrowband beam-pattern for delay-and-sum beamformer with linear microphone array

2.3.1. *Linear Microphone Array*

We consider a linear uniform array of 5 omnidirectional microphones. The aperture size is 38.25 cm and the sampling frequency is 8 kHz. The spacing between the microphones is set to half of the wavelength of a (2246 Hz) frequency in the speech spectrum below which most of the power exists to suppress the majority of the grating lobes. The speech broadband beam-pattern as expressed through equation 6 is plotted in figure 1 vs. the narrowband beam-pattern of a delay-and-sum beamformer for frequencies equal to 250, 500, 1000 and 2246 Hz. As we can observe, there is no null in the speech beam-pattern and the microphone array captures the signal in all directions, whereas there are clearly 8 nulls in the 2246 Hz beam-pattern. The 3 dB beamwidth for speech beam-pattern is about 80° while for the for a 2246Hz beampattern, it is 20°. The difference quantifies the reduction in directivity of the microphone array for speech signal. The broadband as well as the narrowband beam-patterns for a superdirective beamformer are illustrated in figure 2. As we can see, the same argument holds as for the delay-and-sum beamformer; i.e., the speech beam-pattern does not have any null and thus null steering by a uniform linear microphone array is impractical. Side-lobe levels at 0° and 180° are -20 dB.

2.3.2. *Circular Microphone Array*

The circular microphone array is a common structure used for meeting acquisition and robotics. We consider here a scenario of 8 uniformly placed omnidirectional microphones with diameter of 20 cm. So the spacing between the microphones is equal to our previous study with linear microphone array to minimize the majority of the sidelobes. Figure 3 shows the speech beam-pattern vs. the narrowband beam-pattern of the delay-and-sum beamformer at 250, 500, 1000 and 2246 Hz. We observe that the 3 dB beamwidth at 2246 Hz is about 20°. There are 6 nulls and the opposite sidelobe level is -9 dB. However, the speech beam-pattern does not have any null and the 3 dB beamwidth is about 80°; it is evident that the directivity is much smaller. In figure 4, the broadband vs. narrowband beam-patterns are contrasted for a superdirective beamformer. The 3 dB beamwidth at 2246 Hz is about 25° and the opposite sidelobe level is about -8 dB whereas the 3 dB beamwidth for the speech beam-pattern is 40° while the sidelobe level is decreased to -28 dB. The 8 nulls exhibited in the 2246 Hz beampattern do not exist in the speech

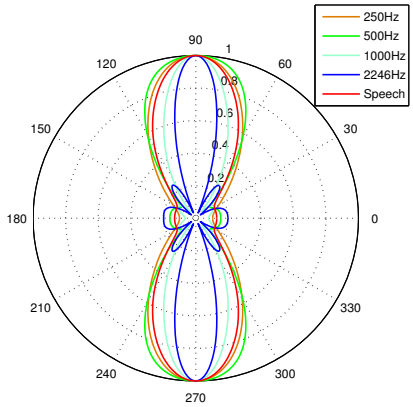

Fig. 2. Speech beam-pattern vs. narrowband beam-pattern for superdirective beamformer with linear microphone array

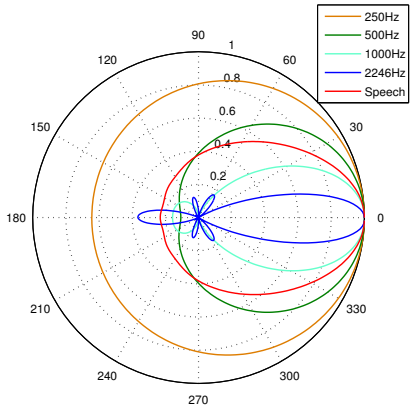

Fig. 3. Speech beam-pattern vs. narrowband beam-pattern for delay-and-sum beamformer with circular array

beam-pattern.

3. EXPERIMENTS

We now present some evaluations on real data recordings to see how the theory formulated in the previous section matches multi-channel speech acquisition. We also demonstrate the relationship between the beam-pattern and distant speech recognition performance.

3.1. Speech Acquisition

The experiments are performed in the framework of Multi-channel Overlapping Numbers Corpus [10]. We used the single speaker recordings captured by an 8-channel circular microphone array of diameter 20 cm. The speaker is located at azimuth and elevation 135° and 25° respectively, related to center of microphone array. The hypothesized room set-up for the theory described in Section 2 is a 6 sided enclosure with reflection coefficients of zero. We also assumed that the mutual coupling between the microphones is almost zero. However, our empirical evaluations are performed in a meeting room fully furnished and the microphone array is installed

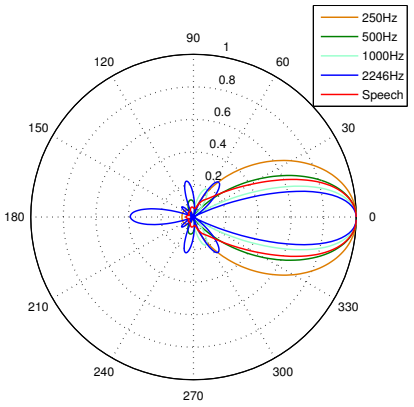

Fig. 4. Speech beampattern vs. narrowband beam-pattern for superdirective beamformer with circular microphone array.

on wood. Hence, the hypothesized ideal condition does not hold in practice and this could justify the difference between the theoretical expected results and our experimental evaluation.
The measured beam-pattern for a superdirective beamformer is illustrated in figure 5(d) which has a similarity to what we obtain in figure 4. It is wider however due to the wall reflections. Because we use a planar array, the beam-pattern has a fan style that captures more signals from the walls. Moreover, the recordings were made in a moderately reverberant 8.2 m $\times$ 3.6 m $\times$ 2.4 m rectangular room. The reverberation time is estimated about 200 ms. The corresponding reflection ratio, β is about 0.8, calculated via Eyring's formula [11]. This can justify the -8 dB opposite sidelobe level exhibited in figure 5(d). The measured beam-pattern of the delay-and-sum beamformer as depicted in 5(e) also looks similar to the speech beampattern that we obtain in figure 3. Similar to the previous argument, we obtain a -8 dB increase in the opposite sidelobe level. In summary, our experiments demonstrate the proposed theory of broadband beam-pattern; however, it does not take reverberation and mutual coupling into account.

3.2. Speech Recognition

The automatic speech recognition (ASR) scenario was designed to broadly mirror that of Moore and McCowan [12]. A typical front-end was constructed using the HTK toolkit with 25 ms frames at a rate of 10 ms. This produced 12 mel-cepstra plus the zeroth coefficient and the first and second time derivatives; 39 features in total. Cepstral Mean Normalization (CMN) is applied to the feature vectors which improves the speech recognition performance about 15%. The average SNR of the recordings is 9 dB. The dominated noise has diffuse characteristics [13] so we use a McCowan post-filter to achieve a higher accuracy using superdirective beamformer. Whereas Moore and McCowan [12] performed MAP adaptation, our results were obtained by training directly on beam-formed data. The maximum ASR accuracy of the system is about 95%. We extract the ASR pattern by scanning all directions using a superdirective beamformer. Figure 5(a) shows the ASR results after normalization with respect to the maximum word recognition rate.

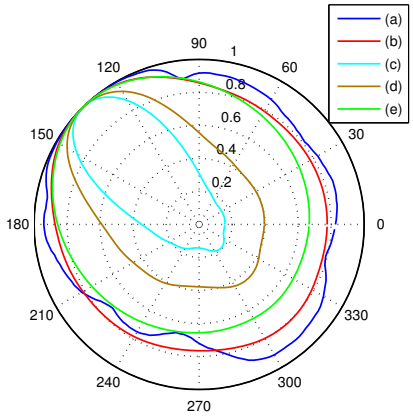

Fig. 5. Beam-pattern, power-pattern and distant speech recognition performance for circular microphone array used in MONC recordings: (a) normalized ASR word accuracy, (b) logarithm of measured speech power-pattern plus one, (c) measured speech power-pattern, (d) measured speech beam-pattern, (a)-(d) are plotted for superdirective beamformer and (e) measured speech beam-pattern of delay-and-sum beamformer

3.3. Discussion

Neither the predicted beam-pattern nor measured power pattern match the ASR pattern; we would not expect an exact match as they are different measures. However, notice that the logarithm of the power-pattern plus one fits the ASR pattern reasonably well. It is illustrated in figure 5 (b).

To investigate the link between the broadband power-pattern and speech recognition performance, we consider a simplistic, but informative view of an ASR front-end. The acoustic signal is treated in Fourier domain and a non-linear frequency warping is applied through a filterbank. To obtain the cepstrum features, a logarithm is applied followed by another linear transform to achieve the decorrelation and dimensionality reduction. The logarithm is motivated to approximate the sensitivity of the ear. The CMN is a common practice to reduce the channel effect and improve the performance. Garner [14] showed that in the presence of CMN, the feature presented to the ASR decoder is (a linear transform of)

$$c = \log(1 + s/n), \tag{10}$$

where s/n is the signal to noise ratio in the spectral domain, and c is the normalized cepstrum. We tentatively conclude that this logarithm of the speech power pattern plus one is a reasonable predictor of ASR performance. This measure is also the Shannon channel capacity for a Gaussian channel, suggesting an information theoretic relationship too.

4. CONCLUSIONS

We described a new method for characterizing a microphone array beam-pattern for the broadband spectrum of a speech signal. We demonstrated the theoretical implications on a variety of microphone array designs, suggesting a generally wider beam-pattern with small sidelobes that can show the response of the microphone array for broadband signals. We also observed that the broadband beam-pattern provides a good estimation of the observable beam in all directions. A high similarly is observed between the logarithm of power-pattern plus one and the speech recognition performance. We hence conclude that the proposed method is a useful approach for analysis of the microphone array structure in terms of speech quality and recognition in hand-free acquisition and it could be further exploited in the design of an optimal geometry for speech applications.

5. ACKNOWLEDGMENTS

This work was supported by the Swiss National Science Foundation under the National Center of Competence in Research (NCCR) on "Interactive Multi-modal Information Management" (IM2) and the European Union 7^{th} Framework Programme IST Integrating Project "Together Anywhere Together Anytime" (TA2, FP7-214793).

6. REFERENCES

[1] M. L. Seltzer, "Microphone array processing for robust speech recognition," in *PhD Thesis, Carnegie Mellon University*, 2001.

[2] J. M. Valin, J. Rouat, and F. Michaud, "Enhanced robot audition based on microphone array source separation with post-filter," in *International Conference on Intelligent Robots and Systems,(IROS)*, 2004.

[3] M. S. Brandstein and H. F. Silverman, "A practical methodology for speech source localization with microphone arrays," in *Computer Speech and Language*, 1997.

[4] B. D. Van Veen and K. M. Buckley, "Beamforming: A versatile approach to spatial filtering," *IEEE Signal Processing Magazing*, vol. 5, 1988.

[5] H. Cox, R. M. Zeskind, and T. Kooij, "Practical supergain," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 34, 1986.

[6] J. O. Coleman and R. J. Vanderbei, "Random-process formulation of computationally efficient performance measures for wideband arrays in the far field," in *Proc. Midwest Symp. on Circuits and Systems (MWSCAS)*, 1999.

[7] M. Wolfel and J. McDonough, *Distant Speech Recognition*, chapter 13, pp. 409–492, John Wiley & Sons Ltd., 2009.

[8] M. Brandstein and D. Ward, *Microphone Arrays*, chapter 2, pp. 19–37, Springer, 2001.

[9] H. L. Van Trees, *Optimum Array Processing*, chapter 2, pp. 17–89, John Wiley & Sons Ltd., 2002.

[10] "The multichannel overlapping numbers corpus," Idiap resources available online:, `http://www.cslu.ogi.edu/corpora/monc.pdf`.

[11] C. F. Eyring, "Reverberation time in dead rooms," *Journal of the Acoustical Society of America*, vol. 1, pp. 217–241, 1930.

[12] D. C. Moore and I. A. McCowan, "Microphone array speech recognition : Experiments on overlapping speech in meeting," in *Proceedings of ICASSP*, 2003.

[13] M. J. Taghizadeh, P. Garner, H. Bourlard, H. R. Abutalebi, and A. Asaei, "An integrated framework for multi-channel multi-source localization and voice activity detection," in *IEEE Workshop on Hands-free Speech Communication and Microphone Arrays (HSCMA)*, 2011.

[14] Philip N. Garner, "Cepstral normalisation and the signal to noise ratio spectrum in automatic speech recognition," *Speech Communication*, vol. 53, no. 8, pp. 991–1001, October 2011.