MODEL-BASED SPARSE COMPONENT ANALYSIS FOR REVERBERANT SPEECH LOCALIZATION

Afsaneh Asaei[1], *Hervé Bourlard*[1,2], *Mohammad J. Taghizadeh*[1,2] *and Volkan Cevher*[2]

[1]Idiap Research Institute, Martigny, Switzerland
[2]Ecole Polytechnique Fédérale de Lausanne (EPFL), Switzerland
{afsaneh.asaei, herve.bourlard, mohammad.taghizadeh}@idiap.ch, volkan.cevher@epfl.ch

ABSTRACT

This paper studies the problem of multiple speaker localization via speech separation based on model-based sparse recovery. We compare and contrast computational sparse optimization methods incorporating harmonicity and block structures as well as autoregressive dependencies underlying spectrographic representation of speech signals. The results demonstrate the effectiveness of block sparse Bayesian learning framework incorporating autoregressive correlations to achieve a highly accurate localization performance. Furthermore, significant improvement is achieved using ad-hoc microphones for data acquisition set-up compared to the compact microphone array.

Index Terms— Structured sparsity, Reverberant speech localization, Autoregressive modeling, Ad-hoc microphone array

1. INTRODUCTION

Speech localization in the clutter of voice and acoustic multipath is an active area of research on microphone arrays for hands-free speech communication. The accurate knowledge of the speaker location is essential for an effective beampattern steering and interference suppression [1, 2]. We briefly review the main approaches to address this problem.

High Resolution Spectral Estimation: These approaches are based on analysis of the received signals' covariance matrix and impose a stationarity assumption for accurate estimation [3]. Important techniques applied for speech localization include minimum variance spectral estimation as well as eigen-analysis methods such as multiple signal classification (MUSIC). The underlying hypotheses are not quite realistic in reverberant speech localization and alternative strategies have been usually considered [4, 5].

Time Difference Of Arrival (TDOA) Estimation: Another approach is based on TDOA estimation of the sources with respect to a pair of sensors. The generalized cross correlation (GCC) is the most common technique for TDOA estimation where the idea is basically to map the peak location of the cross-correlation function of the signal of two microphones to an angular spectrum. A weighting scheme is usually employed to increase the robustness of this approach to noise and multi-path effects. Maximum likelihood estimation of the weights has been considered as an optimal approach in the presence of uncorrelated noise, while the phase transform (PHAT) has been shown to be effective to overcome reverberation ambiguities [6, 7]. In addition to the GCC-PHAT, identification of the speaker-microphone acoustic channel has been incorporated for TDOA estimation and reverberant speech localization [8, 9]. However, despite of being practical and robust, TDOA-based techniques do not offer a high update rate. Alternative strategies have thus been sought for multiple-target tracking and adaptive beam-steering [10, 11].

Beamformer Steered Response Power (SRP): In this approach, the space is scanned by steering a microphone array beam-pattern and finding the direction associated to the maximum power. Delay-and-sum, minimum variance beamformers, and generalized side-lobe canceler have been the most effective methods for speaker localization [12]. The SRP-based approaches have a higher effective update rate compared to TDOA-based methods, and are applicable in multi-party scenarios using phase-transform weighting scheme [13].

In this paper, we adopt our speech separation framework using sparse component analysis [14] and conduct the evaluations in terms of speech localization [15]. We analyze the reverberant mixtures of speech signals in spectro-temporal domain. The planar area of the room is discretized into a dense grid such that the speakers are located at particular cells exclusively. A spatio-spectral sparse representation is obtained by concatenating the spectral components attributed to the sources located on the grid. The compressive acoustic measurements associated to the microphone array recordings are characterized using Image model of multipath propagation. The spatio-spectral sparse representation is estimated from the compressive array measurements using sparse optimization methods where the supports of high energy components indicate the source locations. The computational approaches to model-based sparse recovery of spectrographic speech are compared and contrasted considering block, harmonic as well as autoregressive dependencies.

The rest of the paper is organized as follows: Section 2 explains the premises underlying model-based sparse component analysis of reverberant recordings, and sets up the formulation of reverberant speech source localization. The structured sparsity models underlying speech components are elaborated in Section 3 followed by the computational approaches to model-based sparse recovery in Sections 4. Section 5 presents the details of the experiments. Conclusions are drawn in Section 6. The notations used in this paper are as follows

- ⋄ $g \in \{1, \ldots, G\}$: number of a cell on a grids.
- ⋄ $n \in \{1, \ldots, N\}$: number of source; $N \ll G$.
- ⋄ $m \in \{1, \ldots, M\}$: number of microphones; $M < N$.
- ⋄ $f \in \{1, \ldots, F\}$: number of spectral coefficients.
- ⋄ $\{S, \mathcal{S}\}$: spectral representation of single/all source signals.
- ⋄ $\{X, \mathcal{X}\}$: spectral representation of single/all micro. signals.
- ⋄ Φ: microphone array manifold matrix.

AA and MT are supported by the Swiss National Science Foundation under the National Center of Competence in Research (NCCR) on Interactive Multi-modal Information Management (IM2). VC is supported in part by the European Commission under Grant MIRG-268398, ERC Future Proof and SNF 200021-132548.

References

[1] A. Waibel, M. Bett, F. Metze, K. Ries, T. Schaaf, T. Schultz, H. Soltau, H. Yu, and K. Zechner, "Advances in automatic meeting record creation and access," in *Proc. of ICASSP*, 2001.

[2] A. Asaei, M. J. Taghizadeh, M. Bahrololum, and M. Ghanbari, "Verified speaker localization utilizing voicing level in split-bands," *Signal Processing*, vol. 89(6), 2009.

[3] J. Dmochowski, S. Benesty, and S. Affes, "Broadband music: opportunities and challenges for multiple source localization," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2007.

[4] S. Araki, H. Sawada, R. Mukai, and S. Makino, "DOA estimation for multiple sparse sources with arbitrarily arranged multiple sensors," *Journal of Signal Processing Systems*, vol. 63(2), 2011.

[5] K. Kumatani, J. McDonough, and B. Raj, "Microphone array processing for distant speech recognition: From close-talking microphones to far-field sensors," *IEEE Signal Processing Magazine, Special Issue on Fundamental Technologies in Modern Speech Recognition*, 2012.

[6] M. Omologo and P. Svaizer, "Acoustic source localization in noisy and reverberant environments using CSP analysis," in *Proc. of ICASSP*, 1996.

[7] C. Blandin, A. Ozerov, and E. Vincent, "Multi-source TDOA estimation in reverberant audio using angular spectra and clustering," *Signal Processing*, vol. 92, 2012.

[8] F. Ribeir, C. Zhang, D. Florencio, and D. Ba, "Using reverberation to improve range and elevation discrimination in sound source localization," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18(7), 2010.

[9] S. Nam and R. Gribonval, "Physics-driven structured cosparse modeling for source localization," in *Proc. of ICASSP*, 2012.

[10] M. S. Brandstein and H. F. Silverman, "A practical methodology for speech source localization with microphone arrays," *Computer Speech and Language*, vol. 11(2), 1997.

[11] M. Omologo F. Nesta, P. Svaizer, "Cumulative state coherence transform for a robust two-channel multiple source localization," in *Int. Conf. on Independent Component Analysis and Signal Separation (ICA)*, 2009.

[12] J. H. DiBiase, *A High-Accuracy, Low-Latency Technique for Talker Localization in Reverberant Environments Using Microphone Arrays*, Ph.D. thesis, 1993.

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

[14] A. Asaei, H. Bourlard, and V. Cevher, "Model-based compressive sensing for distant multi-party speech recognition," in *Proc. of ICASSP*, 2011.

[15] Afsaneh Asaei, *Model-based Sparse Component Analysis for Multiparty Distant Speech Recognition*, Ph.D. thesis, Ecole Polytechnique Federal de Lausanne (EPFL), 2013.

[16] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *Journal of Acoustical Society of America*, vol. 60(s1), 1979.

[17] A. Asaei, M. Golbabaee, H. Bourlard, and V. Cevher, "Structured sparsity models for multiparty speech recovery from reverberant recordings," *IEEE Trans. on Speech and Audio Processing (accepted)*, 2013.

[18] R. G. Baraniuk, V. Cevher, M. F. Duarte, and C. Hegde, "Model-based compressive sensing," *IEEE Transactions in Information Theory*, 2010.

[19] J. A. Tropp and S. J. Wright, "Computational methods for sparse solution of linear inverse problems,," *Proceedings of the IEEE*, vol. 98, 2010.

[20] T. Blumensath and M. E. Davies, "Iterative hard thresholding for compressed sensing," *Applied and Computational Harmonic Analysis*, vol. 27(3), pp. 265–274, 2009.

[21] A. Kyrillidis and V. Cevher, "Recipes on hard thresholding methods," in *Proceedings of CAMSAP*, 2011.

[22] E. V. D. Berg and M. P. Friedlander, "Probing the pareto frontier for basis pursuit solutions,," *SIAM Journal on Scientific Computing*, 2008, `http://www.cs.ubc.ca/labs/scl/spgl1`.

[23] D. P. Wipf and B. D. Rao, "An empirical bayesian strategy for solving the simultaneous sparse approximation problem," *IEEE Transactions on Signal Processing*, vol. 55(7), 2007.

[24] Z. Zhang and B. D. Rao, "Extension of SBL algorithms for the recovery of block sparse signals with intra-block correlation," *IEEE Transactions on Signal Processing*, 2012.

[25] D. Wang and G. J. Brown, *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications*, Wiley-IEEE Press, 2006.

[26] T. Blumensath and M. E. Davies, "Gradient pursuits," *IEEE Transactions on Signal Processing*, vol. 56, pp. 2370–2382, 2008.

[27] Z. Zhang and B. D. Rao, "Sparse signal recovery with temporally correlated source vectors using sparse bayesian learning," *IEEE Journal of Selected Topics in Signal Processing*, vol. 5(5), 2011.

[28] M. Lincoln, I. McCowan, I. Vepa, and H. K. Maganti, "The multichannel wall street journal audio visual corpus (mc-wsj-av): Specification and initial experiments," in *IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2005.

[29] P. Boufounos, P. Smaragdis, and B. Raj, "Joint sparsity models for wideband array processing," in *Wavelets and Sparsity XIV, SPIE Optics and Photonics*, 2011.

[30] J. Le Roux, P. T. Boufounos, K. Kang, and J. R. Hershey, "Source localization in reverberant environments using sparse optimization," in *Proc. of ICASSP*, 2013.

[31] A. Asaei, H. Bourlard, and P. N. Garner, "Sparse component analysis for speech recognition in mullti-speaker environment," in *Proceeding of INTERSPEECH*, 2010.

[32] B. Yegnanarayana, C. Avendano, H. Hermansky, and P. S. Murthy, "Speech enhancement using linear prediction residual," *Speech Communication*, vol. 28(1), 1999.

[33] L. Carin, "On the relationship between compressive sensing and random sensor arrays," *IEEE Antennas and Propagation Magazine*, vol. 51, pp. 72–81, 2009.