Max-Planck-Institut für Mathematik in den Naturwissenschaften Leipzig

Learning Binaural Spectrogram Features for Azimuthal Speaker Localization

(revised version: May 2013)

by

Wiktor Młynarski

Preprint no.: 44 2013

Learning Binaural Spectrogram Features for Azimuthal Speaker Localization

Wiktor Młynarski

[1]Max-Planck Institute for Mathematics in the Sciences, Leipzig, Germany

mlynar@mis.mpg.de

Abstract

Spatial localization of speech and other natural sounds with rich spectro-temporal structure is a computationally challenging task. It requires extraction of features which are informative about speaker's position and yet invariant to sound level and spectral modulation present in the signal. This paper demonstrates that this can be achieved with Independent Component Analysis (ICA) applied to binaural speech spectrograms. A small subset of learned Independent Components (ICs) captures signal structure imposed by outer ears. A Gaussian Classifier trained on those features, performs accurate localization on the azimuthal plane. The remaining majority of ICs have position invariant distributions, and can be used to reconstruct the spectrogram of the original sound source.

Index Terms: speech localization, spectrogram, binaural

1. Introduction

Binaural hearing mechanisms exploit between-ear disparities to infer spatial position of the sound source. According to the well known Duplex Theory [1], interaural time differences (ITDs) constitute a major cue for low frequency sound localization and sounds of high frequency (> 1500 Hz) are localized with interaural level differences (ILDs). However in natural hearing conditions, spectro-temporal properties of sounds vary continuously, hence combinations of cues available to the listener also change. The auditory system must extract position invariant information regardless of sound quality, separating "what" and "where" information.

Even though temporal differences on the order of microseconds are of substantial importance for sound localization, binaural neurons in the higher areas of the auditory pathway can be characterized with Spectro Temporal Receptive Fields (STRFS), where the temporal resolution is much more coarse - of the order of milliseconds. Despite such loss of temporal accuracy, many of those neurons reveal sharp position selectivity [2]. Inspired by their properties, present work uses speech sounds represented as log-spectrograms similarly as during receptive field estimation of auditory neurons [3]

As its main result, this paper demonstrates that speech localization can be accomplished with a Gaussian Classifier trained on features learned with Independent Component Analysis (ICA). ICA can be viewed as a generative model, which maximizes independence between latent variables within the limits of the linear transformation. In consequence, features (or basis functions) learned with ICA capture correlations between all data dimensions and in this way identify "causes" i.e. coherent structures present in the dataset. The algorithm learns spectral interaural differences , resulting from convolution with Head Related Impulse Responses (HRIR), filters characterizing sound distortion in outer ears. Features learned with ICA capture variability present in the data due to sound's identity and spatial position into two separate feature subsets. It maybe therefore thought of as a simple model of "what" and "where" information separation in the real auditory system.

2. Data and algorithms

The overview of the data generation process, feature learning and classification is presented on Figure 1. The dataset consisting of binaural spectrograms of randomly sampled speech segments convolved with HRIRs was preprocessed with PCA. Independent components were learned from the PCA whitened data and a Gaussian classifier was trained with activations of a subset of selected components.

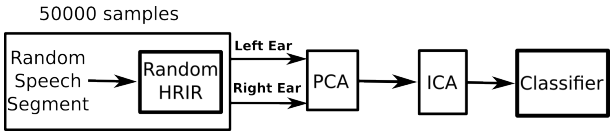

Figure 1: *Data generation and learning architecture*

2.0.1. Speech sounds

A speech corpus from Handbook of International Phonetic Association [4] was used. It includes human narratives in 27 languages sampled at 20 kHz. All sounds were down-sampled to 16 kHz and bandpass-filtered between 200 Hz and 6 kHz.

2.0.2. Head Related Impulse Responses

The binaural signal was obtained by convolving speech sounds with HRIRs from the LISTEN database [5]. The database contains HRIRs from 51 human subjects measured with 15 deg azimuthal resolution. Filters of a randomly selected subject were used for simulating binaural sounds.

2.0.3. Training data generation

Training dataset was created by sampling 50000 speech segments 216 ms long each. Every segment was convolved with an HRIR corresponding to a position on the azimuthal plane, randomly drawn from the set of 24 availible ones. In this way, speech quality and its spatial position were independent factors in the data. A spectrogram was computed on resulting left and right ear sounds. The waveforms were divided into 25 overlapping 16 ms intervals. A squared Fourier Transform was performed on each interval, at 256 frequencies spaced logarithmically between 200 and 4000 Hz using Goertzel algorithm. Resulting spectrogram was transformed with a logarithmic function. A similar signal representation was used by Carlson et al. in their recent work [6] in attempt to model early stages of pro-

cessing in the auditory system. Left and right ear spectrograms were concatenated and subjected to further processing (see Section 3).

3. Independent Component Analysis

Independent Component Analysis (ICA) is a family of linear data transformations, which attempt to minimize statistical dependence between the learned features [7]. Removing dependencies between data dimensions or "redundancy reduction" is a hypothesized general principle of nervous system functioning [8]. A number of studies have investigated redundancy reduction in the auditory system using experimental [9] and computational [6, 10, 11, 12] approaches. In [6, 10, 11], the authors learned representations of natural sounds with ICA or a similar learning algorithm to show that learned features resemble receptive fields in the mammalian auditory system. In a more functional study [12], the authors demonstrate that non-redundant, overcomplete representation of natural sounds learned with Non-negative Matrix Factorization (an ICA-like algorithm) may be used to perform monaural sound localization.

Let us assume that X is a data matrix with columns $x(t)$ corresponding to data samples of dimensionality equal to the number of rows n. ICA seeks a linear transformation of the data i.e. matrix W (called the filter matrix) which minimizes dependence between data dimensions in the transformed space. The ICA transformation can be therefore formulated as:

$$WX = S \tag{1}$$

where S is a matrix of independent component activations. Rows of S correspond to data dimensions and its columns $s(t)$ to data samples represented in the independent component space. ICA can be also viewed as a linear generative model of the data, as defined by equations 2 3

$$p(s(t)) = \prod_{i}^{n} p(s_i(t)) \tag{2}$$

$$p(x(t)|s(t), A) = \delta(x(t) - As(t)) \tag{3}$$

In the equations above $x(t)$ denotes the $t-th$ data sample, $s(t)$ a representation of that sample in the independent component space, and $A = W^{-1}$ is a matrix, columns of which constitute independent components (ICs) (interchangeably named also basis functions or features through the rest of the paper). Marginal distributions of latent coefficients are usually assumed to be sparse i.e. of positive kurtosis. Typically, a logistic distribution is chosen [7] and this distribution was also used in the present work. ICs were learned with a gradient ascent on the data log-likelihood function [7]. The data dimensionality for each ear was equal to the number of time intervals times number of frequencies hence total dimensionality was $25 \times 256 \times 2 = 12800$. Prior to the IC learning, the dimensionality of the data was reduced with PCA from 12800 to 324 dimensions which preserved 99.4% of variance.

4. Feature selection and position classification

The primary goal of this study was to find a low-dimensional representation of speech-spectrograms allowing for an accurate speaker localization. Since a discrete set of 24 spatial positions was used in the simulation, the localization task can be posed as a classification problem. Given representation of a binaural speech spectrogram in a learned feature basis, the task is to assign the sample with a class label C_{est} representing the spatial position. In order to perform classification a Gaussian classifier (GC) was trained on the learned features. GC models the marginal distribution of latent coefficients used for classificationas a mixture of Gaussian distributions s_g such that:

$$p(s_g) = \sum_{C} p(s_g|C)p(C) \tag{4}$$

$$p(s_g|C) = \mathcal{N}(\mu_C, D_C) \tag{5}$$

where μ_C, D_C are class-specific mean vector and covariance matrix. Since the prior on class labels is uniform, classification can be formulated as a maximum-likelihood estimation.

$$C_{est} = \arg\max_{C} p(s_g|C) \tag{6}$$

The resulting procedure iterates over all class labels and returns the one maximizing the probability of observed data sample. Learned independent components were sorted according to correlation between their monaural parts (see section 5.2). The classifier was trained firstly on the first feature from that list, then on first and second and so on. Position-informative features were selected from the entire feature set as those, which influenced the classifier performance and decreased the classification error measured as an average difference in degrees between the decoded position C_{est} and the actual position C_{act} (see sections 5.2, 5.3).

5. Results

Since position-related head filtering and speech content were drawn independently, the initial hypothesis was that ICA would learn separate basis functions capturing spatial information and speech structure separately. A linear transformation should be sufficient to do so, since convolution becomes equivalent to addition of filter and the signal in the log-spectrogram domain. The following sections demonstrate that it was indeed, the case by analyzing structure of learned basis functions, properties of coefficient distributions and showing classification results.

5.1. Properties of learned features

The entire set of learned basis functions was sorted according to the similarity between their left and right ear parts, measured with a correlation coefficient. In the next step, the entire IC population was divided into two subsets - those with binaural correlation smaller than and equal to 1. The first subset consisted of 10 ICs with different monaural parts, depicted on Figure 2 A – they are called "binaural features" through the remaining part of the paper. They captured mostly patterns of interauraly negatively correlated signal i.e. if signal in the one ear had a positive value, it was negative in the other ear. Most binaural features were temporally stable, and revealed no temporal modulation, except for features numbered 5 and 6 in Figure 2 A. Because sound segments had stable spatial positions, the temporal stability of binaural features suggests that they represent position-specific spectral information resulting from HRIR filtering. It is important to notice, that binaural features were not exactly deconvolved HRIRs. Remaining, binaurally correlated

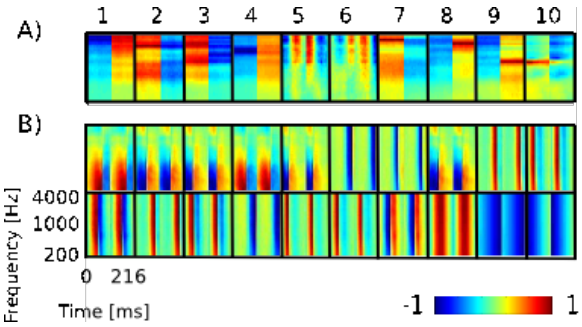

Figure 2: *Examples of Independent components learned on binaural speech spectrograms. Left and right ear parts are separated with a short black line. A) A subset of ten components with dissimilar left and right ear parts. B) Twenty out of* 314 *components with the same left and right ear parts.*

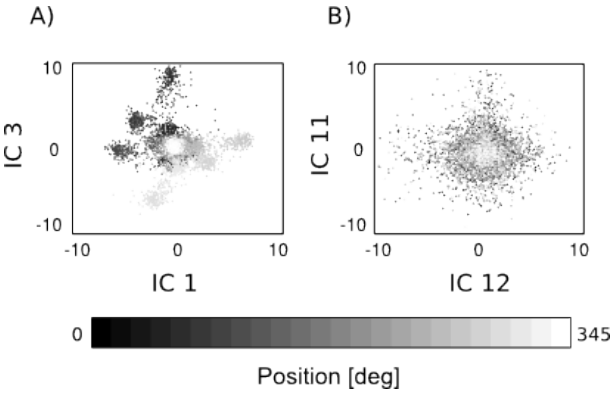

Figure 3: *Dependencies between component coefficients. Scatter plots of component coefficients inferred for* 5000 *different speech samples. A) Two binaurally dissimilar components (IC* 1 *and IC* 3*) - a clear clustering pattern is visible. B) Two monaural components (IC* 11 *and IC* 12*), coefficient distributions do not vary depending on sound position.*

ICs formed a much larger set. An example of 20 out of 314 is depicted in Figure 2 B. They were called "monaural", since the left and right ear parts were essentially the same, and no interaural differences were present. Monaural features captured the spectrotemporal structure of the speech signal, such as onset, offsets, harmonics stacks and many others. They resemble a sparse code of speech spectrograms used in a recent study [6] which argued that receptive fields of neurons in the Inferior Colliculus are adapted to the statistics of natural sounds.

5.2. Position specific coefficient distributions

Marginal histograms of linear coefficients conformed to the model assumptions, i.e. they fit a logistic distribution well and revealed no obvious dependencies. Conditioning on a speaker position though, uncovered additional, useful structure in the data. Figure 3 A depicts a scatter plot of activations of two binaural features – number 1 and 3. Each point corresponds to a single speech sample, and its color to a spatial position it was played from. Coefficients remain close to zero for most of the time (which results from the sparsity assumption), and the global shape of the point cloud approximates a two-dimensional logistic distribution. Most importantly however, a strong clustering is visible – sounds originating from the same position, were represented with very similar feature values. Such property characterized all temporally stable binaural features (i.e. numbered from $1-4$ and from $7-10$).

Figure 3 B presents a scatter plot of signal projections on two monaural features. The global structure is similar as in the Figure 3 A. The substantial difference is that coefficient distributions are stationary that is they do not change, depending on the sound position. Visually, one can not differentiate any clusters, which indicates that monaural features do not carry position-specific information.

5.3. Classification and reconstruction

To quantify contributions of each learned feature to position identification, a Gaussian Classifier was used. It was trained on a progressive number of features added consecutively from a population sorted by binaural correlation. For training, 50 per cent (25000) of the data samples was used and the remaining 50 per cent was used for cross-validation.

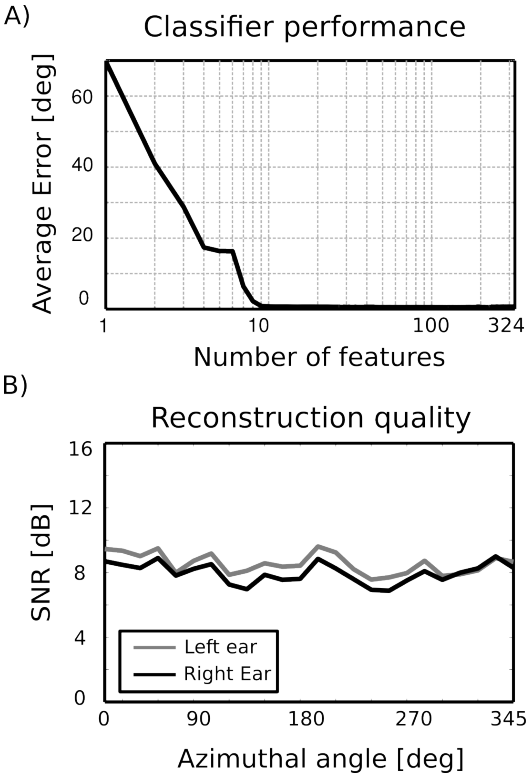

Figure 4: *Classification and reconstruction performance. A) Average classification error as a function of a number of used features (ICs). B) Average reconstruction quality of an original sound spectrogram using monaural features only.*

Classification error was measured in degrees, as an average circular distance between the classifier output and the actual sound position. The classification results as a function of number of used features are presented in Figure 4 A (please note that feature numbers are on a log-scale). It is clearly visible that first 10 features i.e. all binaural ones saturate the classifiers performance by decreasing the classification error to zero. Further adding of monaural features does not influence the error. Interestingly, temporally modulated binaural basis functions – num-

bers $5, 6$ do not contribute to the localization accuracy, which is visible as a small plateau on the plot.

The remaining part of the feature set represents the speech structure. It can be therefore used for the reconstruction of the original, single channel sound spectrogram, prior to the convolution with an HRIR. Since left and right ear parts of monaural components are the same, and therefore redundant, a single one of them can be used to perform such reconstruction. Quality of reconstruction was measured in dB as signal to noise ratio (SNR). Figure 4 B presents SNR as a function of sound position. Gray and black lines represent reconstruction using right and left ear parts of independent components respectively.

The SNR varies very weakly with sound position remaining mostly at a constant level of 10 dB and does not seem to depend on the ear used. This position invariant reconstruction quality indicates that learned representation indeed separates features imposed by ear filtering from the sound itself.

6. Discussion and conclusions

This paper shows that linear features of binaural spectrograms learned with ICA suffice to reliably decode the azimuthal position of the speaker. The binaural, position-discriminative features are essentially complex combinations of level differences across different channels. This result can be interesting from a neuroscientific point of view, showing a possible mechanism, by which a biological auditory system can achieve sound identity invariance and extract spatial information from a binaural signal. From a machine learning perspective, the dictionary learned with ICA is particularly interesting since it contains both: discriminative and reconstructive features, and it is known that learning of both at the same time requires application of different learning algorithms in a general case [13,14]. Further work is needed to explore applications of the presented system to online speaker localization and extraction of other, more complex spatial aspects, such as sound source motion or features useful in solving the cocktail-party problem.

7. Acknowledgements

The author would like to thank Nils Bertschinger for helpful discussions. This work was funded by DFG Graduate College "InterNeuro".

8. References

[1] B. Grothe, M. Pecka and D. McAlpine. "Mechanisms of sound localization in mammals." Physiological Reviews 90, no. 3, pp: 983-1012, 2010

[2] J.W.H. Schnupp, T.D. Mrsic-Flogel and A.J. King. "Linear processing of spatial cues in primary auditory cortex." Nature 414, no. 6860, pp: 200-204, 2001.

[3] P. Gill et al. "Sound representation methods for spectro-temporal receptive field estimation."Journal of computational neuroscience 1, no. 1, pp: 5-20, 2006.

[4] International Phonetic Association. "Handbook of the International Phonetic Association: A Guide to the Use of the International Phonetic Alphabet." Cambridge University Press, 1999

[5] O. Warusfel et al., LISTEN HRTF Database `http://recherche.ircam.fr/equipes/salles/listen/index.html`

[6] N.L. Carlson, V.L. Ming and M.R. DeWeese "Sparse Codes for Speech Predict Spectrotemporal Receptive Fields in the Inferior Colliculus" PLoS Computational Biology 8.7, 2012

[7] A. Hyvrinen, J. Karhunen and E. Oja. Independent Component Analysis., John Wiley & Sons, Inc., 2002.

[8] H.B. Barlow. "Possible principles underlying the transformation of sensory messages." Sensory communication, 1961

[9] G. Chechik et al. "Reduction of information redundancy in the ascending auditory pathway. "Neuron" 51, no. 3, pp: 359-368, 2006

[10] K.P. Kording, P. Konig and D.J. Klein. "Learning of sparse auditory receptive fields." In Neural Networks, 2002. IJCNN'02. Proceedings of the 2002 International Joint Conference on, vol. 2, pp. 1103-1108. IEEE, 2002.

[11] M.S. Lewicki "Efficient coding of natural sounds." Nature Neuroscience 5, no. 4, pp: 356-363, 2002

[12] H. Asari, B.A. Pearlmutter and A.M. Zador, "Sparse representations for the cocktail party problem. The Journal of neuroscience 26, no. 28, pp: 7477-7490, 2006

[13] I. Tošić and P. Frossard. "Dictionary learning." Signal Processing Magazine, IEEE 28, no. 2, pp: 27-38, 2011

[14] K. Huang and S. Aviyente. "Sparse representation for signal classification."Advances in neural information processing systems 19, p: 609, 2007