Monaural Azimuth Localization Using Spectral Dynamics of Speech

Roi Kliper[1], *Hendrik Kayser*[2], *Daphna Weinshall*[1], *Israel Nelken*[1] *and Jörn Anemüller*[2]

[1]Interdisciplinary Center for Neural Computation, Hebrew University of Jerusalem, Israel
[2]Dept. of Physics, Carl von Ossietzky University, Oldenburg, Germany
kliper@cs.huji.ac.il, hendrik.kayser@uni-oldenburg.de

Abstract

We tackle the task of localizing speech signals on the horizontal plane using monaural cues. We show that monaural cues as incorporated in speech are efficiently captured by amplitude modulation spectra patterns. We demonstrate that by using these patterns, a linear Support Vector Machine can use directionality-related information to learn to discriminate and classify sound location at high resolution. We propose a straightforward and robust way of integrating information from two ears. Each ear is treated as an independent processor and information is integrated at the decision level thus resolving, to a large extent, ambiguity in location.

Index Terms: Speech localization, Amplitude modulation, Monaural Processing

1. Introduction

The ability to perform speech localization is an important prerequisite for daily human communication. Accurate speech localization is a fundamental building block for advanced speech processing that handles problems such as stream segregation, source separation, source enhancement and denoising, ultimately producing increased speech intelligibility.

A sound wave generated by an external source is diffracted by the head and external ear (as well as other objects in the environment). The resulting changes in the temporal and intensity characteristics of the acoustical stimuli provide cues about the locus of the sound relative to the head. These localization cues have traditionally been divided into *Binaural cues* and *Monaural cues* and various studies have explored their relative efficacy in determining sound localization. Recent efforts have focused on understanding the integration of these different types of information, these efforts have spread from physiological research through psychoacoustic research to application derived research.

This paper employs Amplitude Modulation Spectra (AMS) patterns (see Section 2.3) as representation of speech signals to perform monaural localization. While the efficacy of this representation for speech recognition has already been demonstrated, we show that these features can simultaneously capture location information included in a non-specific speech signal due to direction-dependent filtering by the human head and pinna. We show that the existence of this information allows monaural localization. We propose and demonstrate the idea of treating the two ears as two parallel processors, each processing monaural information and reaching a hypothesis about the location of a given source. Integration between the two ears is achieved, in this view, at the level of the decision rather then at the level of analysis.

1.1. Binaural localization

Binaural hearing is a well established source of information for the localization of sound sources in space, it builds upon two distinctive properties of incoming sounds: interaural time differences (ITDs) and interaural level differences (ILDs). These differences arise from the fact that the two ears are separated by both space and an acoustically opaque mass (the head). Models of the processing of these cues build upon carefully constructed coincidence detectors, and are supported by both physiological and psychoacoustic findings.

While such models for sound localization are both elegant and robust, binaural cues are limited in their effective frequency band [1] and present a major challenge to the integrative capabilities of the nervous system requiring very high accuracy in highly structured labeled lines. Furthermore, to the extent that the head and ears are symmetrical, interaural differences should provide no cue to the vertical location of a sound source on the median plane nor would they contribute to the resolution of the front to back confusion. While acknowledging these limitations, artificial system designers have favored models inspired by the binaural models for sound localization. Some challenges to the primacy of binaural localization cues have indeed been made (e.g. [2]); however, the role of monaural cues in sound localization is, to a great extent, only brought into focus in situations where binaural differences are nonexistent.

1.2. Monaural localization

For a single ear, the changes in temporal and intensity characteristics of the signal are generally described by head related impulse responses (HRIR), a generalization of which are Monaural Room Impulse Reposes (MRIR). This parsimonious description captures all the directional influence a signal may have received on its way to the ear drum. For example, to some abstraction the influence of the pinna is to produce multiple paths to the ear canal, among them a direct path and a reflection from the structure of the pinna. The addition of a direct path with a delayed path of the same signal produces a comb-filtered spectrum containing a characteristic structure of peaks and notches. The pinna thus acts as a direction-dependent filter which strongly affects the HRIR at high frequencies.

Several attempts to explain and exploit monaural cues for sound localization have been made. Generally speaking, these studies rely on the differential way in which the HRIR affects the spectrum of the input signal and require some interaction and/or comparison between different spectral bands of the signal. In this respect, these studies compel to the fact that the signal appearing at the eardrum has no reference point but itself. A second requirement for monaural localization is some statistical model of the source signal or equivalently some restriction to its statistical properties; directional sensitivity in itself cannot be exploited for localization of a general signal. If no assumptions on the nature of the signal are made, such a system is underdetermined and may result in ambiguous localization as each sound signal can be manipulated such that it is perceived as coming from all other locations. Assuming that the source to be localized is the statistically restricted set of speech signals is a step towards resolving this ambiguity. In monaural localization of sound, and more specifically in monaural localization of speech one should bear in mind that high intelligibility and accurate localization often represent competing requirements where intelligibility requires minimum distortion while localization requires direction dependent distortions.

1.3. Physiological and psychoacoustic findings

Psychoacoustic experiments have demonstrated monaural localization of a sound source on both the horizontal and the vertical planes [1]. In a recent paper Shub et al [3] have demonstrated the ability of human subjects to monaurally discriminate and classify different directions. Our experiments follow their paradigm and show that by employing a simple machine learning approach these results can be reproduced and surpassed (see Section 3.2).

Chase & Young [4] explored how different acoustic localization cues are coded in the inferior Colliculus (IC). The rationale was that the IC integrates binaural cues (ITD, ILD) and monaural cues (spectral information). Their results suggest that different cues converge to different degrees in different neurons: ITD and ILD are coded most strongly by spike rate while the spectral envelope of the signal is coded by the temporal pattern of the spikes. Localization in the vertical plane is often thought to be a purely monaural ability but recent psychophysical studies [5] have shown that both ears are used to determine the elevation, with the relative contribution of each ear varying with horizontal location. Our binaural experiment successfully implements this idea for localization in the horizontal plane (see Section 3.2).

2. Experimental setting

2.1. Data

Experiments were carried out using a well known speech database (see Section 2.1.1) and a database of HRIRs (see Section 2.1.2), these together provide a rich and reproducible setting with complete control over the experimental conditions.

2.1.1. Speech

Speech data was taken from the TIMIT Speech corpus [6] which provides recordings of 630 different speakers each reading ten phonetically rich sentences recorded at a sampling rate of 16 kHz. The segmentation into train and test data was adopted from the corpus. Both of the sets were further divided into direction specific subsets sampled randomly from the dataset and concatenated. No intersection between any of the subsets was allowed, resulting in 155 s train and 57 s test data for each of 37 directions.

2.1.2. Head-related impulse responses

A database of head-related impulse responses [7] was employed to introduce directional characteristics to the speech data. The database provides, among others, HRIRs (or MRIRs, respectively) measured on an head and torso simulator equipped with in-ear microphones under anechoic conditions. The azimuthal resolution is 5° covering the full azimuthal plane. MRIRs from an elevation angle of 0° (no elevation) and a radius of 3 m were taken. The initial delay contained in the MRIRs was removed and the remaining impulse response was cut to a length of 10 ms.

The audio data was convolved with the MRIR corresponding to a specific direction, resulting in monaural sound signals containing the input to either the left or the right ear. All data was scaled to unit variance to compensate for overall level differences between signals according to different directions. For experiments carried out in the presence of diffuse noise, pink random noise was generated artificially. The noise signal was convolved with each direction's MRIR from the full circle except for the one the target speech source was impinging from, thus approximating an isotropic noise field. After convolution, the noise was added to the directional speech source with the desired SNR. The coordinate system for the azimuth angles is relative to the center of the horizontal plane and is 0° in front of the head -90° and 90° in front of the left and right ear.

2.2. Classification

Linear support vector machines (SVM [8]) were employed to conduct training of discriminative models using AMS features. The task consists of either binary classification, where a model was trained to discriminate between the directions of two speech sources, or multi-class classification, where a set of models was trained to estimate the absolute direction of an impinging speech source in a *1 vs. all* approach given a set of more than two directions.

2.3. AMS features and their extraction

Following the structure of the temporal envelope has shown to be crucial in human and machine recognition of speech. As a consequence, low modulation frequencies were employed for several tasks in speech processing and acoustic scene analysis. Such features deliver a robust and generalized representation of speech signals and are known to be largely invariant to speaker and channel variations such as pitch and spectral distortions in the input signal. They showed to be a robust representation of speech even in challenging conditions, e.g. in speech detection experiments [9]. The features employed here are based on AMS [10] and are calculated as follows (cf. Figure 1):

The amplitude modulation spectrogram analyzes sound signals with respect to their modulation content by decomposing them into time, frequency and modulation-frequency components. It is computed by first extracting the spectral envelopes of the acoustic signal via a short-term Fourier Transformation

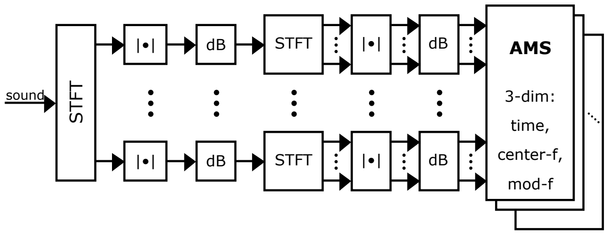

Figure 1: ***AMS calculation flow:*** *Extraction of amplitude modulation features generates a 3-dimensional representation of the input. For each* 100 *ms signal segment a* 256×9 *pattern is extracted, overlap of neighboring patterns is* 50 *ms.*

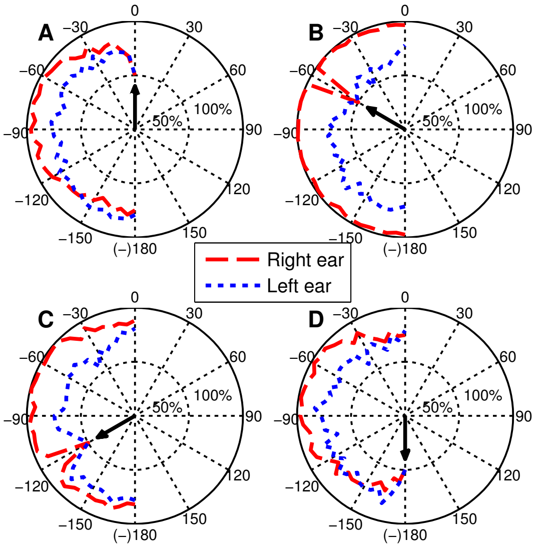

Figure 2: ***Binary classification:*** *The accuracies gained by the right (contralateral) ear are shown in dashed red and by the left (ipsilateral) ear are shown in dotted blue. Four target signals where classified against a reference signal drawn one at a time from the left hemisphere at a 5° resolution. Performance in terms of testing accuracy is shown on the radial axis. Position of reference signals are shown on the circular axis. Target signal is pointed by a black arrow.*

(STFT) with a 32 ms Hamming window with a shift of 2 ms followed by computation of the log-energy. To extract modulation energy in each spectral band, another STFT is applied employing a 100 ms Hamming window with a shift of 50 ms.

Finally, the log-energy is computed again and the DC-component of the modulation spectrum, containing the acoustic frequency spectrum of the input signal, and the first acoustic frequency component of the resulting AMS pattern are removed to disregard spectral properties of the signal. Modulation frequencies above 100 Hz are also removed. The resulting AMS spectra for each time frame span 256 frequency bands from $32 - 8000$ Hz and modulation frequency bands covering the range from 20 Hz to 100 Hz with a resolution of 10 Hz. The employed range patterns lie above modulation frequencies used in speech recognition and detection (typically ≤ 16 Hz).

3. Experiments and results

In the following we first present experiments of a discrimination task which is a popular task in psychoacoustics used to evaluate minimum audible difference. We then show results of multi-direction classification experiments under different noise conditions. Finally, results of a straightforward *winner takes all* formalism of binaural integration are presented.

3.1. Minimum audible angle difference

In this experiment a set of models for binary classification was trained to discriminate between two sound sources s_1 and s_2 impinging from different directions. s_1 (target) was located at positions from 0° to -180° in steps of -60° while the angular position of s_2 (reference) is varied with a resolution of 5° on the same half circle. The results are shown in Figure 2.

The superiority of the contralateral monaural cues over the ipsilateral can be read from the consistent improved accuracy of the right ear on the left hemisphere (red line above blue line in e.g. Figure 2 B). Contralateral cues achieve a nearly perfect performance discriminating the target from references around it at a resolution of 5°. This makes sense as heavier distortion has been introduced in the contralateral case thus allowing better directional sensitivity. We note that, as we normalized the energy at the eardrum, our results cannot be explained away using level differences (which will appear in real scenarios). One can also note (1) The appearance of moderate front to back confusion in the cases where s_1 is located exactly in front or behind the listener (see Figure 2 A & D). (2) Increased performance in the frontal half space (compare Figure 2 B & C). Similar discrimination characteristics can be found in psychoacoustic monaural localization experiments carried out in [3]; however, although their experiment was done with stimuli ten times longer, our results surpass theirs.

3.2. Localization of sound sources in the presence of noise and the integration of both ears

In the multi-class classification experiment we first evaluated the influence of averaging the features on classification performance. For that, a sliding average over time was applied to the AMS patterns before models were trained. Averaging was conducted with a sliding window of length ranging from 0 s (no averaging) to 5 s in steps of 0.25 s. Training and testing was conducted without noise (infinite SNR). Furthermore robustness against diffuse pink noise was investigated: noise was introduced as described in Section 2.1 with an SNR varying from 10 dB to 60 dB in steps of 10 dB.

Figure 3 displays the results of this experiment for 12 directions distributed equally spaced over the complete horizontal plane from -180° to 150°. The accuracies gained by the left and the right ear independently in addition to the the performance achieved by the integration of information from both ears are shown. For each ear, one model was trained and tested on one condition given by the SNR and length of averaging. Different noise conditions appear in Figure 3 A-G and show a clear dependency on noise level ranging from guessing probability (8.3%) at SNR 10 dB to 56.5% and 56.2% for the single ears and 76.5% for the combination of to ears where no noise is present (inf). Average accuracy dependent on the averaging length is read against the x-axis in each figure and shows a global maximum at around 1 s suggesting that longer averaging windows are erasing localization information. Integration of the information from both ears was done in a *winner takes all* manner where the more confident ear was taken as the reliable one. Confidence was assumed to be positively correlated to the margin from the model's separating hyperplane. Integration of the two ears achieves up to 20% absolute boost in accuracy.

Besides the overall average testing accuracy the confusion between different directions is a significant figure of merit. The confusion matrices obtained for an averaging length of 1 s in the clean condition from the left and the right ear and from the integration of both ears are shown in Figure 4. Clear preference for the contralateral ear can be read in Figure 4 A & B. The ipsilateral ear regains significant classification capabilities only when the angular distance is larger than 60° (allowing directional cues to come into effect). Integration of the information from the two ears solves most of the confusion, leaving a slightly increased confusion around the center supposably due to noise in the confidence of the classifiers.

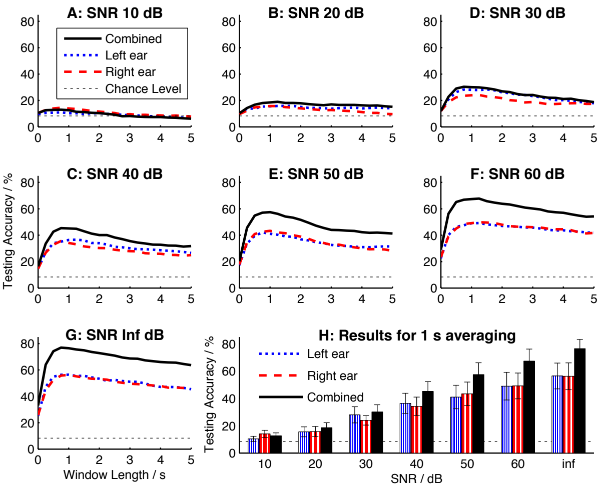

Figure 3: ***Multi-class classification at different SNRs:*** *A-G: Performance of the localization on* 12 *directions for different SNRs and averaging times of the features. Blue and red lines denote the accuracy achieved by the single ears, black the performance of the integrated approach. H: Average and standard error of accuracy for each SNR.*

4. Discussion and conclusions

We have demonstrated that amplitude modulation spectra patterns in modulation frequencies above those commonly used for other speech applications are efficient in monaural speech localization. Our results suggest that under clean and moderate noise conditions, accurate speech localization can be achieved using the information obtained by a single ear, without distorting intelligibility related information. The experiments showed a maximal performance for an integration time of around 1 s, which corresponds to the choice of parameters in monaural localization experiments conducted by Shub et al. Keeping in mind that the analysis of the stimuli was done using a rather technical approach, further investigation incorporating auditory models as a preprocessing stage is a natural step. While monaural localization was proven to be feasible, the integration of information processed parallel in the two ears is highly beneficial as each ear performs better on the contralateral hemisphere. Relating to physiological research we hypothesize that the IC may host such an integration mechanism.

This success in speech source localization highlights the potential use of the described method in the context of other speech processing applications such as source separation, signal to noise ratio estimation, and noise reduction. Drawing to the design of artificial systems, the implementation of such an approach in ear-worn hearing assistive systems is an interesting application as it does not necessarily demand for a technically costly cross-linking in the case of a bilateral means.

5. Acknowledgements

Support of the *Gatsby Charitable Foundation* is acknowledged by R. Kliper. Support by the DFG Collaborative Research Centre SFB/TRR 31 *The Active Auditory System* acknowledged by H. Kayser and J. Anemüller.

6. References

[1] J. Middlebrooks and D. Green, “Sound localization by human listeners,” *Annual Review of Psychology*, vol. 42, no. 1, pp. 135–159, 1991.

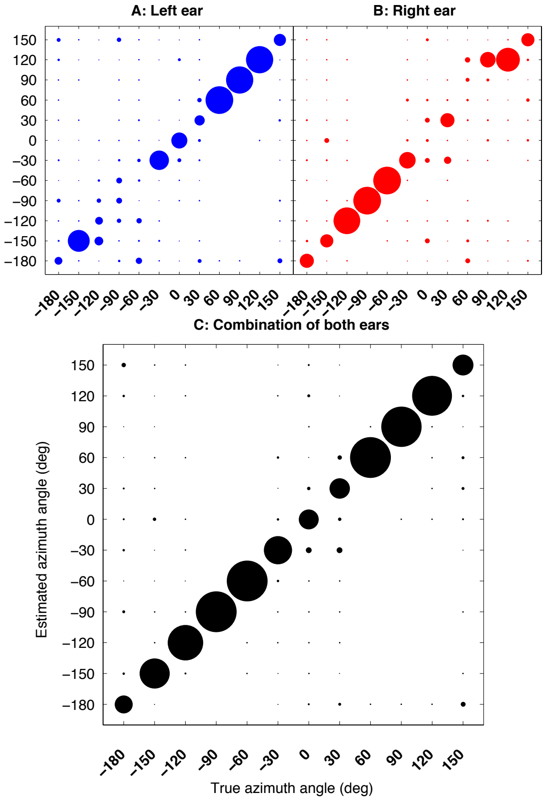

Figure 4: ***Confusion matrix for multi-direction classification:*** *Confusion matrices for single ears (upper row) and after integration of both ears for the clean condition and averaging window length of* 1 *s. Coincidence of target and reference is given by the radius of the circle.*

[2] H. Fisher and S. Freedman, “The role of the pinna in auditory localization.” *Journal of Auditory research*, 1968.

[3] D. Shub, S. Carr, Y. Kong, and H. Colburn, “Discrimination and identification of azimuth using spectral shape,” *The Journal of the Acoustical Society of America*, vol. 124, p. 3132, 2008.

[4] S. Chase and E. Young, “Cues for sound localization are encoded in multiple aspects of spike trains in the inferior colliculus,” *Journal of neurophysiology*, vol. 99, no. 4, p. 1672, 2008.

[5] P. Hofman and A. Van Opstal, “Binaural weighting of pinna cues in human sound localization,” *Experimental brain research*, vol. 148, no. 4, pp. 458–470, 2003.

[6] J. Garofolo, L. Lamel, W. Fisher, J. Fiscus, D. Pallett, and N. Dahlgren, *DARPA TIMIT Acoustic Phonetic Continuous Speech Corpus CDROM*, U.S. Dept. of Commerce NTIS, Gaithersburg, MD, 1990.

[7] H. Kayser, S. Ewert, J. Anemüller, T. Rohdenburg, V. Hohmann, and B. Kollmeier, “Database of multichannel in-ear and behind-the-ear head-related and binaural roomimpulse responses,” *EURASIP Journal on Advances in Signal Processing*, p. 10, 2009.

[8] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin, “LIBLINEAR: A library for large linear classification,” *Journal of Machine Learning Research*, vol. 9, pp. 1871–1874, 2008.

[9] J.-H. Bach, B. Kollmeier, and J. Anemüller, “Modulation-based detection of speech in real background noise: Generalization to novel background classes,” in *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on.* IEEE, 2010, pp. 41–44.

[10] B. Kollmeier and R. Koch, “Speech enhancement based on physiological and psychoacoustical models of modulation perception and binaural interaction,” *The Journal of the Acoustical Society of America*, vol. 95, p. 1593, 1994.