PROCEEDINGS of the
23rd International Congress on Acoustics
9 to 13 September 2019 in Aachen, Germany

Binaural sound localisation directly from the raw waveform

Ning Ma, Paolo Vecchiotti, and Guy J. Brown

Department of Computer Science, University of Sheffield, Sheffield S1 4DP, UK
{n.ma, p.vecchiotti, g.j.brown}@sheffield.ac.uk

Abstract

It is well known that binaural sound localisation is largely based on cues such as the interaural time difference (ITD), or the related interaural phase difference (IPD), and the interaural level difference (ILD). These cues encode a location in terms of the difference between the left and right ears in both arrival time and magnitude. In addition, the outer ear, together with the head, shoulders and torso, form direction-selective filters, whose resonances impose direction-specific patterns into monaural spectral cues. Conventional machine systems for sound localisation attempt to explicitly extract ITD and ILD features. An alternative approach is to estimate the location of a source directly from the waveform arriving at each ear. In this study, we propose a novel binaural machine system that robustly localises a speech source by implicitly extract features from the waveform that are suitable for localisation.

Our approach differs in two important respects from previous studies. First, instead of using explicitly extracted IPDs and ILDs as features, the proposed system uses a convolutional neural network (CNN) framework with 2-dimensional (2-D) kernels that operate directly on the phase spectrum and the magnitude spectrum of the left and right ear signals. The convolution operations between the left and right ears extract IPD and ILD-like features, but have better robustness, particularly in reverberant environments. Secondly, the 2-D convolution kernels also operate along the frequency dimension, and thus allow binaural information and monaural spectral information to be combined effectively within the same neural network architecture. The correlation-based features extracted from both the phase spectra and the magnitude spectra are then concatenated and used as input to fully connected neural network layers in order to map them to the corresponding source elevation. Our evaluation shows that the proposed system is able to accurately estimate the location of a speech source, even in challenging reverberant conditions, and substantially improves upon the performance of previous approaches.

Keywords: Binaural sound localisation, end-to-end, convolutional neural networks, raw waveform.