GROUP DELAY FEATURES FOR SOUND EVENT DETECTION AND LOCALIZATION (TASK 3) OF THE DCASE 2019 CHALLENGE

Technical Report

Eike J. Nustede and Jörn Anemüller

Computational Audition Group
and Cluster of Excellence Hearing4All
Dept. Medical Physics and Acoustics
Carl von Ossietzky University Oldenburg

ABSTRACT

Sound event localization algorithms utilize features that encode a source's time-difference of arrival across an array of microphones. Direct encoding of a signal's phase in subbands is a common representation that is not without its shortcomings, since phase as a circular variable is prone to 2π-wrapping and systematic phase-advancement across frequencies. Group delay encoding may constitute a more robust feature for data-driven algorithms as it represents time-delays of the signal's spectral-band envelopes. Computed through derivation of phase across frequency, it is in practice characterized by a lower degree of variability, resulting in reduced wrapping and to some extent permitting the computation of average group delay across (e.g., Mel-scaled) bands. The present contribution incorporates group delay features into the baseline system of DCASE 2019 task3, supplementing them with amplitude features. System setup is based on the provided baseline system's convolutional recurrent neural network architecure with some variation of its topology.

Index Terms— Audio classification, acoustic source localization, convolutional recurrent neural network, group delay

1. FEATURE EXTRACTION

1.1. Spectral decomposition

Audio input was constitutet by the provided four-channel first-order ambisonic recordings, with one minute long recordings at sampling rate 48000 Hz. The development set consisted of 400 recordings and the evaluation set of 100 recordings [1]. A (complex-valued) spectrogram was computed for each input channel using a short-term Fourier transformation (STFT) with Hamming window of length 2048 and 50% overlap. Only the 1024 positive bins, excluding the zeroth bin, of the hermitian spectrum were used for a total spectrogram dimensionality of (3000, 1024, 4) representing frames, frequency-bands and channels, respectively. Magnitude and phase delay components (see below) were extracted from the spectrograms, and their Mel-band averages were used to supplement the linear-frequency representation. The Mel-band filterbank employed here contained 128 triangularly shaped filters whose center frequencies are linearly spaced on the approximately logarithmic Mel-scale given by

$$m = 2595 \cdot \log_{10}\left(1 + \frac{f}{700}\right), \tag{1}$$

where m denotes Mel-scaled frequency and f linear (Fourier) frequency. Total feature length of 1152 spectral bands results from the concatenation of 1024 Fourier bands with 128 Mel bands, performed separately for each channel's magnitude and group delay representation. Fig. 1 shows an overview of the feature extraction pipeline. Incorporation of concatenated linear and Mel-scaled features for each channel and magnitude plus group delay results in a final feature tensor dimensionality of (3000, 1152, 8) for each audio file.

1.2. Group Delay features

Group delays may serve as features that indicate spatial source position since they may provide a representation that is more robust and supports computations that are non-trivially performed with the acoustic phase. In particular in upper spectral bands, phase may behave semi-randomly in acoustic recordings since a temporal delay induces a phase shift proportional to frequency. Group delays, in turn, represent phase of the signal's envelope and are constant in case of a linear phase system. In practice, they permit to some extent the computation of averages across spectral bands and, thus, a more robust characterization of a sources spatial features. The use of group delay and variations thereof for acoustic event detection has been explored in, e.g., [2, 3]. Group delay is incorporated into the feature extraction pipeline through computation of $\tau_g(k)$ as

$$\tau_g(k) = -\arg\left(\phi(k+1) \cdot \phi^*(k)\right)/\Delta, \tag{2}$$

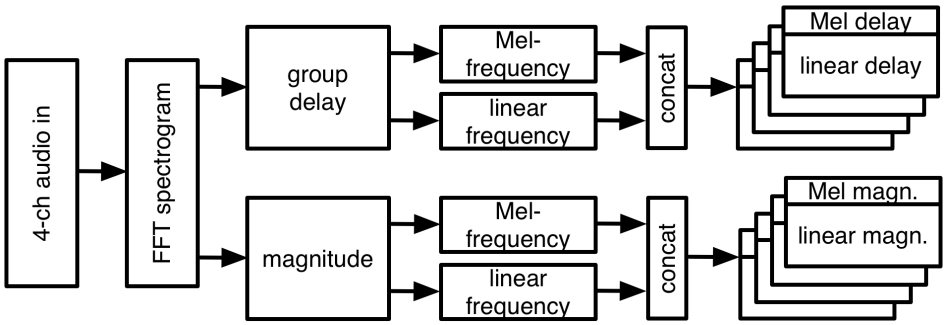

Figure 1: Diagram of proposed feature extraction with group delay in linear and Mel-scaled spectral bands.

where k denotes spectral band index, $\phi(k)$ (complex-valued) STFT, ϕ^* complex conjugate, Δ a spectral resolution parameter, and $\arg(\cdot)$ the argument (polar angle) operator. Spectral bands from 0 Hz to the Nyquist limit (1025 in total), thus, result in 1024 polar coordinate values for the group delay representation, and, subsequent to averaging in Mel bands as indicated above, in 128 Mel-scaled group delay features. They replace the STFT phase spectra of the baseline system, cf. Fig. 1 for an overview.

	Net 1	Net 2	Baseline
#conv. layer	4	4	3
#RNN layers	2	3	2
conv. kernels	3x3	3x3	3x3
pooling layers	1x(8,8,4,2)	1x(8,8,4,2)	1x(8,8,4)
CNN output	(64,128,2)	(64,128,2)	(64,128,4)
total no. param.	552,761	750,137	613,537

Table 1: Summary of model architecture parameters.

2. MODEL ARCHITECTURE

Focusing on exploring different, suitable features, we opted to use the proposed baseline system architecture described in [4]. The CRNN was expanded by one additional convolution layer (Net 1 and Net 2), and one additional recurrent layer (Net 2). Each of the 4 convolution layers consisted of $P = 64$ filters with a max-pooling on the feature axis of size (8, 8, 4, 2). The Output of the CNN was of shape (64, 128, 2) which was then reshaped to (128, 128) being denoted as T x 2 x P in the original work. This vector was then given into the bidirectional RNN with $Q = 128$ GRU nodes. The RNN output, given as T x Q being (128, 128), was fed to two different fully connected networks (FCNs). The first FCN was responsible for the SED task, giving a continuous output between [0, 1] in the shape of (128, 11) corresponding to the 11 sound classes present in this sequence. The second FCN outputs DOA estimates of size (128, 22) with range [-1, 1] for each axis of the sound class location. Cf. Table 1 for an overview of model parameters in the architectures evaluated here.

	Net 1	Net 2	Baseline
DOA angle error	**26.38**	28.88	28.21
DOA frame recall	**86.40**	85.81	84.90
SED error rate	0.3255	**0.2970**	0.3620
SED f-score	80.80	**82.15**	79.40
Overall SELD score	0.2001	**0.1944**	0.2190

Table 2: Performance on development set, average across four folds.

	Net 1
DOA angle error	20.69
DOA frame recall	90.11
SED error rate	0.2197
SED f-score	87.61
Overall SELD score	0.1389

Table 3: Performance on validation set with evaluation setup.

3. RESULTS

Utilizing the proposed features, both networks (Net 1 and Net 2) result in a small improvement of sound event localization and detection (SELD) score of about 10% in relative terms. While network Net 1 shows an improvement in localization performance (DOA error improvement of about 2° and improvement in DOA frame recall), Net 2 produces only a very small improvement in DOA frame recall compared to the baseline system. Increasing the depth of the RNN (Net 2)

provides a minute improvement in SELD Score due to an overall gain in SED metrics, while producing no increase in localization performance. Comparison of network performance on the development set is shown in Table 2, while performance of Net 1 on the evaluation set is displayed in Table 3.

4. CONCLUSION

Based on the results from using group delays as features instead of the phase spectrum, it might be possible to further improve upon DOA estimation using these features. The improvement shown here was obtained with only minor system optimization, which might be the reason for the advantage over baseline being small and moderately consistent. Further research may prove as to which degree group delay features are beneficial to the SELD Task in an optimized architecture.

5. ACKNOWLEDGMENT

Supported by Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) grants SFB 1330/B3 no. 352015383 and FOR 1732.

6. REFERENCES

[1] S. Adavanne, A. Politis, and T. Virtanen, "A multi-room reverberant dataset for sound event localization and detection," in *Submitted to Detection and Classification of Acoustic Scenes and Events 2019 Workshop (DCASE2019)*, 2019. [Online]. Available: https://arxiv.org/abs/1905.08546

[2] R. M. Hegde, H. A. Murthy, and V. R. R. Gadde, "Significance of the modified group delay feature in speech recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 1, pp. 190–202, 2007.

[3] A. Diment, E. Cakir, T. Heittola, and T. Virtanen, "Automatic recognition of environmental sound events using all-pole group delay features," *2015 23rd European Signal Processing Conference (EUSIPCO)*, pp. 729–733, 2015.

[4] S. Adavanne, A. Politis, J. Nikunen, and T. Virtanen, "Sound event localization and detection of overlapping sources using convolutional recurrent neural networks," *IEEE Journal of Selected Topics in Signal Processing*, pp. 34–48, 2018.