Acoustic-based 3D Human Pose Estimation Robust to Human Position
Supplementary Materials

Paper ID: 135

Contents

1. Overview of the Supplementary Materials

This supplementary material provides additional details and discussions about our robust acoustic-based 3D pose estimation independent of the human positions. Please also refer to the qualitative evaluation video, 0135_video.mp4. References related to the main paper are highlighted in blue, while those pertinent to this supplementary material are highlighted in red.

2. Acoustic Feature Extraction

For acoustic feature extraction, we utilize the log-Mel spectrum and Intensity Vector [1].

log-Mel Spectrum is a commonly used feature in acoustic signal analysis. Human hearing is highly sensitive to low-frequency sounds and less sensitive to high-frequency sounds. To accommodate this auditory characteristic, the Mel scale is employed, which converts frequency in Hertz [Hz] to Mel scale [mel] using a Mel filter bank. The Mel filter bank comprises triangular filters arranged in such a way that the resolution is higher at lower frequencies and lower at higher frequencies. This setting allows for finer attention to low-frequency sounds and broader attention to high-frequency sounds. The log-Mel spectrum is derived by first performing a short-time Fourier transform (STFT) on the acoustic sequence s_t, then converting it to the Mel scale.

$$I^{\mathrm{mel}}(k,t) = H_{\mathrm{mel}}(k,f) \cdot \mathcal{F}(s_t(f)) \tag{1}$$

where H_{mel} represents the Mel filter bank, k is the index of the Mel filter bank, and $\mathcal{F}$ denotes the STFT. The resulting $I^{\mathrm{mel}}(k,t)$ is then converted to log scale to produce the feature I^{logmel}. The log-Mel spectrum is calculated for each channel, with the number of channels being four.

Intensity Vector is a feature used in the direction of arrival estimation tasks [1]. It involves performing an STFT on the three acoustic signals along the x, y, and z axes, and a non-directional w axis, totaling four components. Intensity Vector is derived by taking the real part of the product of the conjugate complex of the sound's intensity, $W(f,t)$, and the intensities of sound in each direction, represented by $X(f,t)$, $Y(f,t)$, and $Z(f,t)$.

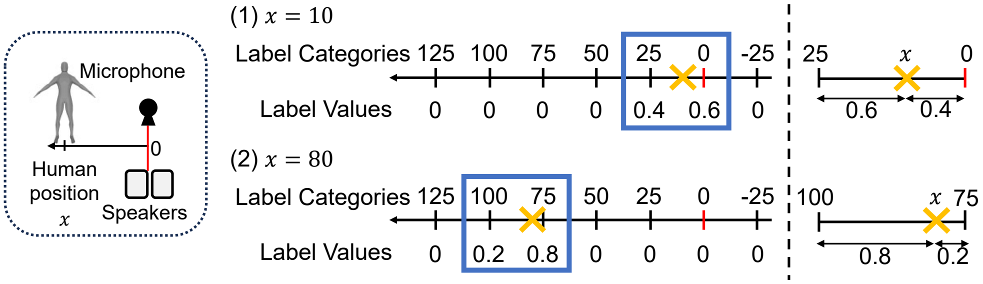
Microphone
0
Human position
x
Speakers
(1) x = 10
Label Categories 125 100 75 50 25 0 -25
Label Values 0 0 0 0 0.4 0.6 0
25 x 0
0.6 0.4
(2) x = 80
Label Categories 125 100 75 50 25 0 -25
Label Values 0 0.2 0.8 0 0 0 0
100 x 75
0.8 0.2

Figure 1. Example of soft label

$$\hat{I}(f,t) = \mathcal{R}\left\{ W^*(f,t) \cdot \begin{pmatrix} X(f,t) \\ Y(f,t) \\ Z(f,t) \end{pmatrix} \right\} \tag{2}$$

$\mathcal{R}\{\cdot\}$ denotes the real part and $*$ represents the complex conjugate. For integration with the log-Mel spectrum as acoustic features in pose estimation module, $\hat{I}'(f,t)$ is transformed into the Mel scale.

$$\hat{I}'(f,t) = H_{\text{mel}}(k,f)\frac{\hat{I}(f,t)}{\|\hat{I}(f,t)\|} \tag{3}$$

Here, $\|\cdot\|$ denotes the L1 norm, and k represents the index of the Mel bins. The Intensity Vector is calculated based on the w-axis, resulting in three channels.

3. Evaluation Metrics

In our main paper, we utilize three evaluation metrics for quantitative comparison: Root Mean Squared Error (**RMSE**), Mean Absolute Error (**MAE**), and Percentage of Correct Keypoints (**PCK**). RMSE and MAE are defined by the Eq. 4 and the Eq. 5:

$$\text{RMSE} = \sqrt{\frac{1}{TJ}\Sigma_{t=1}^{T}\Sigma_{j=1}^{J}(x_t^j - \hat{x}_t^j)^2}, \tag{4}$$

$$\text{MAE} = \frac{1}{TJ}\Sigma_{t=1}^{T}\Sigma_{j=1}^{J}|x_t^j - \hat{x}_t^j|, \tag{5}$$

Here, $\hat{x}_t^j$ represents the coordinate of the j-th joint in the predicted pose at frame t and x_t^j is the ground truth. T is the number of pose frames and J is the total number of joint in a pose.

4. Soft Label for Position Discriminator Module

As discussed in Sec. 3.2, the Position Discriminator Module is implemented as a classifier to categorize the standing positions of the subject. This section explains how we set the ground truth labels used for training the Position Discriminator Module. The labels corresponding to the distance from a line connecting the speaker and microphone are denoted as $[d_1, d_2, \ldots, d_M]$, where M represents the number of label categories. The labels are positioned at equal intervals, satisfying the equation $W = d_{m+1} - d_m$. To annotate each position of the subjects, we consider the position x of the subject as the internal division point between two adjacent labels. Then, we represent x as an internal division point using the two consecutive labels on both sides of x. Generalizing the above concept, the soft label of the subject's position being at d_m and d_{m+1} are given by w and $1-w$ respectively. Furthermore, the probabilities for all labels other than d_m and d_{m+1} are set to zero.

In our main paper, we set $d_1 = -25$, $W = 25$, and $M = 7$, creating label categories $[-25, 0, 25, 50, 75, 100, 125]$. As shown in Fig. 1, for a ground truth position $x = 10$, the label values are given by $[0, 0.6, 0.4, 0, 0, 0, 0]$ $(0 \leq x \leq 25, x = 0.6 \times 0 + 0.4 \times 25)$. Similarly, for a ground truth position $x = 80$, the label values are given by $[0, 0, 0, 0, 0.8, 0.2, 0]$ $(75 \leq x \leq 100, x = 0.8 \times 75 + 0.2 \times 100)$.

Table 1. Comparison Based on Changes in Training Data Volume

Method	half training data			original training data			double training data		
	RMSE (↓)	MAE (↓)	PCKh @0.5 (↑)	RMSE (↓)	MAE (↓)	PCKh @0.5 (↑)	RMSE (↓)	MAE (↓)	PCKh @0.5 (↑)
Ours w/o aug	0.69	0.37	0.48	0.58	0.31	0.55	0.51	0.27	0.61
Ours	**0.63**	**0.33**	**0.53**	**0.53**	**0.28**	**0.60**	**0.45**	**0.23**	**0.68**

5. Evaluation Based on Variations in Training Data Volume

This section conducts accuracy comparisons when the number of frames in the training dataset is increased or decreased, to explore the effects of data augmentation. In conducting these experiments, we did not increase the number of subjects. Instead, we augmented the amount of data per subject. Moreover, the same test dataset was used across different training data volumes. Table 1 shows the differences in accuracy when the training data volume is either halved or doubled. The results demonstrate that data augmentation is effective regardless of the training data volume.

Furthermore, we compare the effects of applying data augmentation versus simply doubling the training data volume without augmentation. When data augmentation is applied to the original training dataset, the estimation accuracy achieved is similar to that obtained by doubling the data volume. However, when data augmentation is applied to a halved training dataset, the accuracy, particularly in terms of RMSE, falls below that achieved using the original dataset. This difference is presumed to arise because data augmentation does not increase the variety of poses.

References

[1] Yin Cao, Turab Iqbal, Qiuqiang Kong, M Galindo, Wenwu Wang, and Mark D Plumbley. Two-stage sound event localization and detection using intensity vector and generalized cross-correlation. In *Proc. Detection Classification Acoustic Scenes Events (DCASE) Challange*, 2019. 1