Acoustic Human Action Classification and 3D Human Pose Estimation

Yutaka Kawashima[1,a)] Yuto Shibata[1] Mariko Isogawa[2] Go Irie[2] Akisato Kimura[2]
Yoshimitsu Aoki[1]

Abstract: Most of the existing methods for inferring human behavior such as actions or poses use visible light or wireless signals as a clue to estimate it. However, visible light is easily restricted by poor lighting conditions (, dark rooms, night roads). Wireless signals often have limited use (, highly instrumented patient care areas where electronic devices must remain off). Unlike these methods, we explore how low-level acoustic signals can provide enough clues to estimate human behavior by active acoustic sensing with a single pair of microphones and loudspeakers (see Fig. 1). This is quite a challenging task since sound is much more diffractive than visible light or RF/WiFi that most of existing method use and therefore covers up the shape of objects in a scene. To this end, we introduce a framework that encodes multichannel audio features into human activity classes or 3D human poses. Our framework only requires a minimal active sensing system with a single pair of ambisonics microphone and loudspeakers to enable estimation. Aiming at capturing subtle sound changes to reveal detailed pose information, we explicitly extract phase features from the recorded audio signals together with typical spectrum features, and feed them into our 1D convolutional neural network to learn non-linear mappings from the features to the target. Our experiments suggest that with the use of only low-dimensional acoustic information, our method outperforms baseline methods.

1. Introduction

The ability to capture human behavior such as poses or activities has many potential applications. Over the last decade, many different technologies including conventional cameras [4], [8], transient light [16], radio frequency (RF) or WiFi measurements [20], [28] have been proposed to infer human activity or pose. However, the optical signals are easily occluded and restricted by poor lighting conditions such as a dark room and night road. RF/WiFi signals are also occluded by water or metal. In addition, the use of wireless signals is also often limited, because electronic devices with transmissions of signals must remain off during flights, as well as in hospital rooms with sensitive electronic systems.

Audio signals, which exist everywhere in our world, have the potential to solve these fatal limitations. We can listen to sounds regardless of the lighting conditions, and acoustic signals do not affect electronic systems. If we use ultrasonic wave that is outside of our audible range, we are not even aware of them. Moreover, since the acoustic signals have a much longer wavelength (meter scale) than visible light (nanometer scale) and RF/WiFi signals (centimeter scale), the signals are less occluded.

Some very recent studies have used acoustic signals for a cross-modal analysis with visual information including scene geometry estimation [5], [23], action recognition [9], visual semantic segmentation [15], and even object understanding [24]. Another line of studies use acoustic signals for sensing humans, like active hand gesture monitoring [18]. Papers that are more relevant to ours are approaches which infer human joints by converting human speech or music to gestures [10], [19], [25]. These methods use human speech as a clue of recovering human gestures/motions. However, no methods have been proposed yet to capture whole human 3D poses given only low-level acoustic signals without any environmental sound that links the signal and target of estimation (, speech, music, sound of specific action).

All of these previous studies motivate the following three questions. First, do low-level acoustic signals have enough information to reconstruct whole 3D human poses? Second, what is the small set of hardware setup for the task? And third, which ones lend themselves to effective inference algorithms?

To answer these questions, this paper takes up a new task, *human activity classification and 3D human pose estimation from only low-level acoustic signals*. This is a quite challenging task. Wavelength of acoustic signals is much longer than optical or RF/WiFi signals. While it could be advantageous for occlusion issues, a longer wavelength is diffractive, making it difficult to distinguish small pose changes. In this work, we explore a solution to this task with minimal equipment configuration using only a single ambisonics microphone as shown in Fig. 1, as opposed to previous methods that use high-definition RGB(D) cameras and RF/WiFi signals from multiple transmitters and receivers. While we do use multiple channels, our microphone is located at a specific single position and has much fewer geometry clues to map

[1] Keio University
[2] Nippon Telegraph and Telephone Corporation
[a)] ykawashima@aoki-medialab.jp

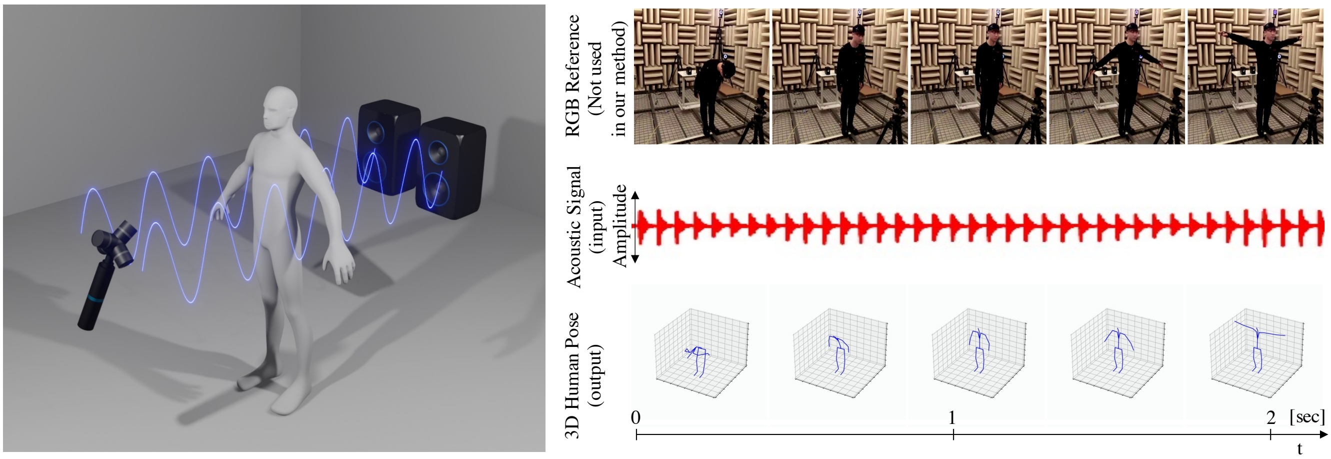


Fig. 1 We propose human action classification and 3D pose estimation given only low-level acoustic signals with a single pair of microphones and loudspeakers. Given an audio feature frame (right-middle), our method estimates frame-by-frame human action labels and 3D poses (right-bottom).

the signals to human activities. Moreover, unlike most previous works that have utilized higher-level semantics such as human speech, music, or a dog barking, our low-level signals do not represent any of this kind of information directly.

To capture human status effectively under such a severe condition, we propose a convolutional neural network (CNN)-based framework designed to employ multi-channel audio features as its inputs and directly output the predicted result of human action classes or 3D body parts joints. If humans occlude acoustic signals emitted from loudspeakers, the subtle “shifts” of arrival time of the incoming acoustic signals occur. Our network model captures these small shifts by explicitly integrate phase features that represent the time difference of arrival (TDOA) and utilizes them to infer human behavior. Since no previous method could tackle this task, there is no public dataset available. Therefore, to train our network, we set up an active acoustic sensing system using a single pair of ambisonics microphones and loudspeakers. Then we actively record the sounds of a time-stretched pulse (TSP) signal emitted from the speaker, synchronized with motion capture (Mocap) data.

To summarize, our contributions are as follows: (1) We are the first to tackle a new task: human activity classification and 3D human pose estimation given only low-level audio signals. (2) We describe a network architecture that directly maps acoustic features to human activity labels and 3D human poses. (3) Since there are no previous methods to carry out this task, we describe how to create a new dataset to train our network model. (4) We provide extensive experimentation and show the effectiveness of our method.

2. Related Work

Table 1 summarizes where our method is positioned among existing approaches that are relevant to ours. Although this paper also describes action recognition as a classification task, Table 1 focuses on the pose estimation task. However, note that our classification task can make a similar argument. The following of this section introduces each work and other relevant approaches in detail.

Human Activity Estimation. Estimating human pose has long been studied by the computer vision community [4], [8]. Although a majority of existing work leverages the fact that the human body is visible from conventional cameras, this line of researches includes a wide variety of solutions in terms of hardware systems and reconstruction algorithms that operate in different parts of the spectrum. Besides visible spectrum (380-740 nm) [16] or near-IR (740-1500 nm) light [6], WiFi and RF (centimeter scale) [20], [28] or even sound waves(meter scale) [18] are used to estimate human behavior.

Operating in a specific part of the spectrum affects the nature of the signal that can be used for human activity or pose estimation. For example, visible signals are easily restricted by poor lighting conditions (, a dark room, night road) and occluded by other objects (, buildings, etc.). RF or WiFi signals enable through-the-wall pose estimation [1], [20], [28] since longer electromagnetic waves tend to pass through objects; however, these signal spectrum are also occluded by some materials (, metal, water, etc.) and are often limited to being used. We have to turn off electronic devices with transmissions of signals during the flights as well as in the hospital rooms with sensitive electronic systems. Audio signals have the potential to overcome such occlusion than other signals due to the longer wavelength. However, this also presents a number of fundamental limitations when estimating human pose: the spatial and angular resolution must be limited, making it hard to distinguish small differences of poses. We in this paper will address these limitations and explore the fundamental potential of active acoustic sensing for action classification and 3D pose estimation.

Acoustic Sensing for Capturing Human Behavior. In this paper, we leveraged active acoustic sensing using a single pair of ambisonics microphones and loudspeakers; hence, our work highly relates to those that leveraged acoustic sensing of some form for capturing human behavior. Passive acoustic sensing has used for gesture recognition [7], [11], [12], on-body sensing [13], activity [9], or even body joints estimation [10], [19], [25]. However, these methods require uses to put wearable devices [7], [11], [12], [13], requires higher level of acoustic information such as

Method	Modality	Occluded by	Required semantics level	Invasiveness
RGB-based [4], [8]	RGB	Any opaque objects	High (image required)	Non-invasive
RF/WiFi-based [1], [20], [26], [28]	RF/WiFi	Metal, water	Low	Non-invasive
Audio to joint [10], [19], [25]	Audio	Soundproof room	High (speech required)	Non-invasive
Audio to hand micro gesture [18]	Audio	Soundproof room	Low	Invasive
Ours	Audio	Soundproof room	Low	Non-invasive

Table 1 Comparisons between existing pose estimation methods and our method.

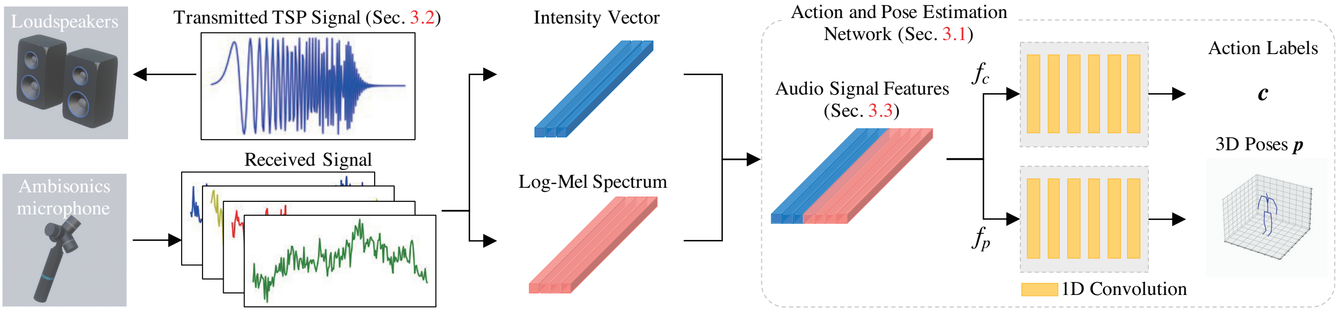


Fig. 2 The overview of our human behavior estimation framework. We transmit TSP signals and receive multichannel audio signals that are encoded into audio signal features. Our method feeds the features into our 1D convolutional neural network to infer human activity class and 3D human poses.

human speech [10], [19], [25], or daily activity sound [9]. Also, these approaches basically reconstruct "gestures", thus are not designed to estimate more "subtle" differences of poses. This is because higher level of acoustic signals often lack enough information to recover finer human poses. Active acoustic sensing has been also researched to monitor a more detailed level of gesture recognition [18], [22], [27]. However, these methods also required to put wearable devices to capture a part of a human body. Inspired by these previous works, we tackle capturing 3D human poses in a non-invasive manner, given acoustic signals only.

3. Methodology

3.1 Acoustic Human Behavior Estimation

Given only a sequence $\mathbf{s} = [s_1, s_2, ..., s_T]$ of raw audio signals, this paper tackles two types of tasks to infer human behavior: (1) inferring detailed human activity class $\mathbf{c} = [c_1, c_2, ..., c_T]$, and (2) estimating 3D human poses $\mathbf{p} = [p_1, p_2, ..., p_T]$, where c_t and p_t represents the action labels and joint 3D positions of frame t, respectively.

The overview of our method is illustrated in Fig. 2. Our proposed framework is made up of an audio feature extraction module that encodes raw audio signals into a sequence of acoustic feature vectors $\mathbf{a} = [a_1, a_2, ..., a_T]$, an action prediction network f_c, and human 3D pose estimation network f_p. Both f_c and f_p are consisted of a 1D CNN with six convolution layers and two fully connected (FC) layers to output $c = f_c(a)$ or $p = f_p(a)$. f_c and f_p are trained separately, depending on which task to choose. With the variable θ that contains all trainable parameters, the training objective for our classification task takes softmax cross-entropy loss $\mathcal{L}_{class}$, while our pose estimation task uses MSE loss $\mathcal{L}_{pose}(\theta) = \frac{1}{T}\Sigma_i^T(\hat{p}_i - p_i)^2$, Here, $\mathcal{L}$ denotes loss function, T indicates the number of samples, $\hat{p}$ represent ground-truth of pose. The following subsections describe the active acoustic sensing in Sec. 3.2 and audio features that are fed into f_c and f_p in Sec. 3.3.

3.2 Active Acoustic Sensing

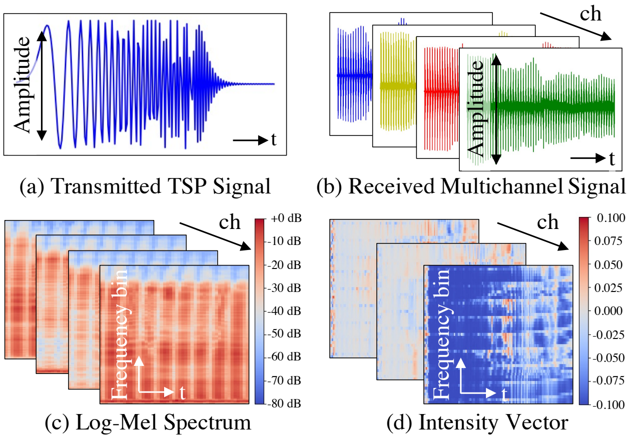
Suppose we have a known sound source and a microphone. The sound emitted from the source bounces off objects in the space and reaches the microphone. Hence, the recorded signal reflects information about the structure of the scene and the position and shape of the objects in the scene. The information we want is the change made to the original sound generated from the source until it is captured by the microphone, which is equivalent to the problem of identifying the room impulse response (RIR), the system transfer function of the environment. Since measuring the RIRs for any given state in advance is impossible, we estimate them using our network in an active acoustic sensing manner.

Following succeeded existing active acoustic sensing [18], we transmit a modulated acoustic signal and pre-process the received signal to emulate RIR. This is a similar approach with "chirp signal" generally applied to FMCW radar which transmits linear sweep frequency-modulated signals. We specifically use time stretched pulse (TSP) as our sound $s'(t)$, which is a kind of swept sine waves designed for RIR measurements.

$$s'(k) = \begin{cases} \exp(\frac{-4\pi jmk^2}{N^2}) \ (0 \leq k \leq \frac{N}{2}), \\ s'^{*}(N-k) \ (\frac{N}{2} < k < N), \end{cases} \qquad (1)$$

where N is the entire waveform length (number of samples), m is a parameter that determines the pulse length of the TSP, k is a parameter that determines a frequency, and superscript * represents a complex conjugate. The inverted TSP signal is defined as a complex conjugate of the TSP signal in a frequency range. For measurement, $s'(k)$ is subjected to inverted Fourier conversion and thereby converted into a signal that takes time as a parameter. The converted signal is reproduced and used. In our system, we emitted TSP signal with the sampling rate 48 kHz.

To effectively capture the 3D structure of the scene, we used a single ambisonics microphone, which consists of four microphones. Each acoustic signal was synchronized and exported as a b-format that has four channels of signals representing a different



(a) Transmitted TSP Signal (b) Received Multichannel Signal

(c) Log-Mel Spectrum (d) Intensity Vector

Fig. 3 Signals and features. (a) The transmitted signal is a time stretched pulse (TSP) signal. (b) The multichannel received signals are transformed as (c) log Mel spectrum and (d) intensity vector, which are concatenated to be fed into our network.

microphone polar pattern, pointing in a specific direction.

3.3 Audio Signal Features

To generate the sequence of the audio feature vectors $\mathbf{a} = [a_1, a_2, ..., a_T]$ as input to our network, one straightforward way would be to directly feed raw signals into a CNN as attempted before in [2] for audio feature learning. However, the multichannel audio that we use is far richer than the monaural sound assumed in their work; hence, it is more important to extract the information needed to make learning stable. Therefore, we extract the (i) intensity vector $I^{\mathrm{intensity}}$ that has $b \times 3$ dimension, including three channels of (x, y, z)-directional components, and (ii) the log Mel spectrum I^{logmel} with $b \times 4$ channels that are often used for sound source localization and audio event detection, respectively. Here, b denotes the number of frequency bins. Since the range of each signal is different between the intensity vector and log Mel spectrum, we standardized them before concatenation. The same standardization is applied at the validation and test time. The final $\mathbf{a}$ is then computed to be a $b \times 7$ tensor. We use $b = 128$ in our implementation.

Intensity Vector. The acoustic signal $s(t)$ that we capture includes four channels: w, x, y, and z. These four-channels of signals include omni-directional, , XYZ-directional components. The instantaneous sound intensity vector can be expressed as $\hat{I} = pv$, where p is the sound pressure obtained from w and $v = (v_x, v_y, v_z)^T$ is the particle velocity vector obtained from x, y, and z. This intensity vector represents the acoustical energy direction of a sound wave. Hence, it can be used to estimate the direction of arrival (DoA) of the sound source, which would be a clue to perceive the scene geometry. In order to concatenate the intensity vector and the log Mel spectrum that we describe later, following [3], we compute intensity vector in the short-time Fourier transform (STFT) domain and the Mel space as follows:

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

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

where W, X, Y, Z are the STFT domain of w, x, y, z, respectively. $\mathcal{R}\{\cdot\}$ indicates the real part, * denotes the conjugate, k is the index of the mel bins, H_{mel} is the mel-bank filter, and $||\cdot||$ represents $L1$ norm. We then standardize $\hat{I}(k,t)$ as follows to extract final intensity vector I that we input into the network:

$$I_i^{\mathrm{intensity}} = \frac{\hat{I}'_i - \overline{\hat{I}'}}{\alpha}, \quad \alpha = \sqrt{\frac{1}{T}\sum_{i=1}^{T}(\hat{I}'_i - \overline{\hat{I}'})^2} \qquad (4)$$

where $I_i^{\mathrm{intensity}}, \hat{I}'_i$ denote ith frame of intensity vector feature after/before standardization and T represents the number of data samples.

Log Mel Spectrum represents an acoustic time-frequency representation and is known for its better performance as the input of convolutional neural network. The Fast Fourier Transform is performed over the received audio signal $s(t)$, and we convert it to the Mel scale as follow:

$$I^{\mathrm{mel}}(k,t) = H_{\mathrm{mel}}(k,f) \cdot \mathcal{F}(s(f,t)), \qquad (5)$$

where k is the index of the Mel bins and H_{mel} represents the Mel-bank filter, and $\mathcal{F}$ is the Fourier transform operation. We then convert it to log scale, and standardize the feature as follows:

$$\hat{I}^{\mathrm{logmel}} = ln(I^{\mathrm{mel}}) \qquad (6)$$

$$I_i^{\mathrm{logmel}} = \frac{\hat{I}_i^{\mathrm{logmel}} - \overline{\hat{I}^{\mathrm{logmel}}}}{\alpha}, \qquad (7)$$

$$\alpha = \sqrt{\frac{1}{T}\sum_{i=1}^{T}(\hat{I}_i^{\mathrm{logmel}} - \overline{\hat{I}^{\mathrm{logmel}}})^2}, \qquad (8)$$

where I_i^{logmel} and $\hat{I}_i^{\mathrm{logmel}}$ denote i-th frame of log Mel spectrum after/before standardization and T represents the number of data samples.

4. Experimental Settings

4.1 Datasets

Motion Capture Suits Dataset. We captured a large set of acoustic measurement data synchronized with Mocap data captured with eight cameras (OptiTrack Prime 17W). As shown in Fig. 4, we used a pair of ambisonics microphone (Zoom H3-VR) and loudspeakers (Sanwa Supply MM-SPU9BK). The acoustic signals were captured in an echoic chamber environment, where the reverberation or other noise can be reduced. The dataset is 1 hour long (equal to 3.6K frames with 10 fps of the frame rate). It consists of eight subjects who were asked to wear Mocap suit and stand between a microphone and loudspeakers while the subjects take action.

In the classification dataset, the subjects took four types of different actions (, standing, sitting, bowing, and even not existing). In the 3D pose estimation dataset, the subjects performed various complex poses: walking, sitting, bending forward, raising both hands, and transitioning between all of these motions. While we labeled each frame in classification datasets following most vision-based methods, note that our 3D pose estimation framework does not require segmenting the pose sequences or labeling them. For pose ground-truth annotation, we used the skeleton of

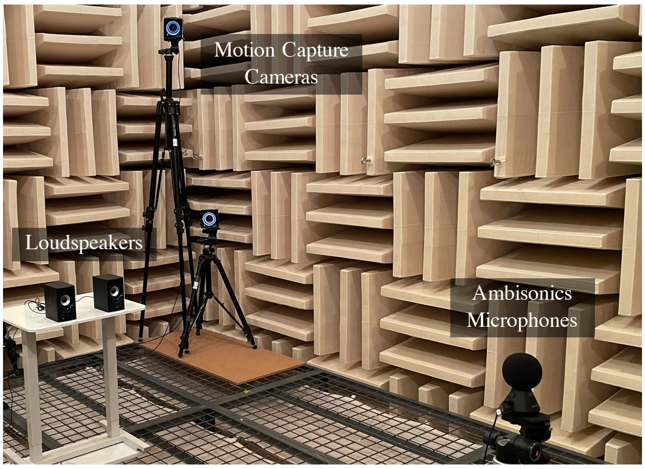


Fig. 4 The setup of our experiments. We used a single pair of loudspeakers and an ambisonics microphone for emitting and receiving audio signals. Subjects were asked to stand between a microphone and loudspeakers while the subjects take action. The ground truth poses were captured via a motion capture system with eight cameras.

21 joints including head, neck, shoulders, arms, forearms, hands, hips, legs, foots, toes, pelvis, and spine. This paper uses this dataset for both training and testing.

In Plain Clothes Dataset. To further showcase our methodâs applicability, we also tested our method on the acoustic signal captured with subjects who do not wear Mocap suits. Although audio signals do pass through normal clothes, the signal attenuation or diffraction may change, depending on what the subjects put on and whether their clothes are tight-fitting. Two subjects were asked to put on plain clothes and act the same as the motion capture suits dataset. Since this dataset does not include ground-truth poses, it is used only for testing purposes.

4.2 Baseline Methods

There is no existing work on 3D full-body human pose estimation from low-level acoustic signals without any environmental sound such as human speeches or music. Therefore, following [10], we compare our network model against the approaches that use audio signals that include music [14], [25]. Both networks were trained with our own dataset for fair architecture comparisons. The detail of each method is as follows:

- **ResNet50-based [14]:** Following [14] that investigates effective CNN-based network architecture for audio classification, we used ResNet50 based baseline.
- **Shlizerman [25]:** We compare our network against one of the state-of-the-art methods for capturing 2D hand and arm poses from passively captured acoustic music [25]. The network employs a unidirectional single layer LSTM and a fully connected layer. To train the network with our own dataset, we modified the input and the last layer of the network so that the network takes our audio feature and outputs 3D poses as regression.

4.3 Evaluation Metrics

For human pose estimation task, we use three types of metrics to evaluate our method: root mean square error (**RMSE**), mean absolute error (**MAE**), and percentage of correct key points (**PCK**). RMSE and MAE measure the average magnitude of the error and the absolute differences between predicted and actual observation of each human joint as:

Table 2 Human action recognition results.

Method	Accuracy(%)
ResNet50-based [14]	88.36
Ours	**89.38**

$$\mathrm{RMSE} = \sqrt{\frac{1}{TJ}\Sigma_{t=1}^{T}\Sigma_{j=1}^{J}(x_t^j - \hat{x}_t^j)^2}, \tag{9}$$

$$\mathrm{MAE} = \frac{1}{TJ}\Sigma_{t=1}^{T}\Sigma_{j=1}^{J}|x_t^j - \hat{x}_t^j|, \tag{10}$$

where x_t^j is the j^{th} joint position of the estimated pose and $\hat{x}_t^j$ is the ground truth. The length of the data and the number of total joints are denoted as T and J, respectively. PCK measures the percentage of the predicted joint locations that are within a specific range from the ground truth. Specifically, this paper applies PCKh@0.5 score that uses a threshold=50% of the head–neck bone link.

4.4 Implementation Details

Audio Signal Features. We used librosa [21] as an audio signal processing library. In the main experiments, we sampled acoustic frames to extract audio features at 30 fps for the classification task, while we used 10 fps for the pose estimation task.

Networks and Training. We use Adam [17] to optimize our network with learning rate 0.003. The network function typically converges after 100 epochs, which takes about an hour for pose estimation and about 10 minutes for pose classification on a GeForce GTX 1080 Ti.

5. Experiments and Results

We conducted five different experiments to investigate our method's efficacy: (1) performance comparison with a baseline method in classification task, (2) comparison with a baseline method in pose estimation task, (3) ablative analysis regarding the audio features, and (4) investigating about the trade-off between the length of time window and estimation accuracy. While these four tests use motion capture suits dataset that includes ground truth, we also conducted (5) qualitative analysis with “in plain clothes” dataset for pose estimation task.

Comparison with Baseline in Action Classification. We first investigated the efficacy of our action recognition network f_c. We trained f_c with the data of randomly selected four subjects and tested the data from two subjects. As shown in the quantitative results in Table 2, we can see our model infers human action classes with 89.38% of estimation accuracy, which outperforms the ResNet50-based baseline by 1.02 points. We consider that our effective audio feature extractor enables the lightweight network with fewer layers when compared with baselines to better fit to capture human actions.

Comparison with Baseline in Pose Estimation. Same as the previous experiment, we use a four-two subjects data split. The quantitative results are shown in Table 3 that shows that our

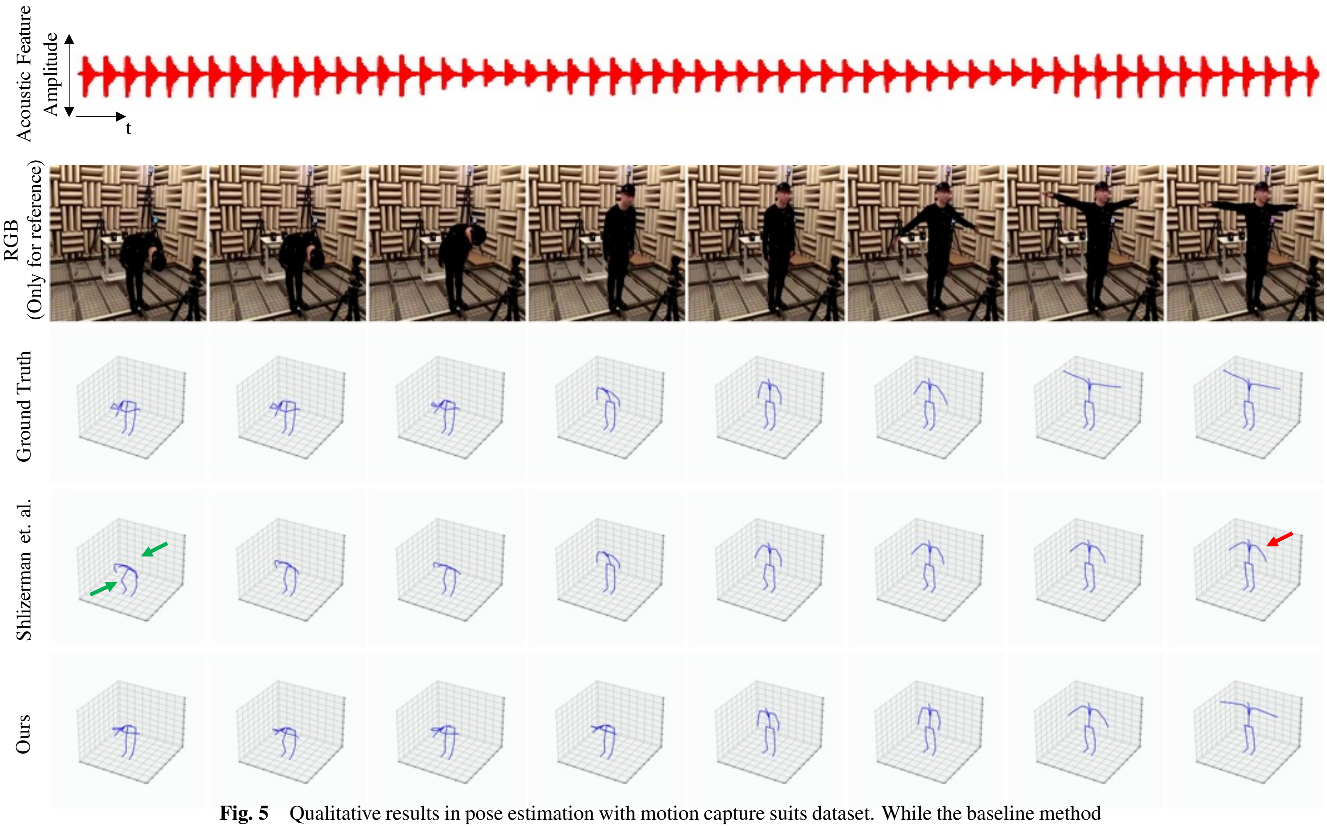


Fig. 5 Qualitative results in pose estimation with motion capture suits dataset. While the baseline method fails to reconstruct finer poses, our method outputs closer poses to ground truth.

Table 3 Human 3D pose estimation results.

Method	RMSE(↓)	MAE(↓)	PCKh@0.5(↑)
Shlizerman [25]	1.17	0.65	0.53
Ours	**0.59**	**0.36**	**0.72**

Table 4 Ablative analysis results.

	Recognition	3D Pose Estimation		
Method	Accuracy (%)	RMSE (↓)	MAE (↓)	PCKh@0.5 (↑)
Ours	**89.38**	**0.59**	**0.36**	**0.72**
w/o $I^{\text{intensity}}$	87.55	0.69	0.42	0.63
w/o I^{logmel}	88.47	0.65	0.39	0.69

Table 5 Investigation on FPS for feature extraction.

Method	RMSE(↓)	MAE(↓)	PCKh@0.5(↑)
Ours (10 fps)	**0.59**	**0.36**	**0.72**
Ours (20 fps)	0.85	0.48	0.60
Ours (30 fps)	1.06	0.58	0.57

method outperforms Shlizerman 's model. We also present qualitative results in Fig. 5. As the green arrows show, Shlizerman 's model was not able to reproduce finer human poses like "bowing" (see the leftmost plot which is outputting the transition-like pose between bowing and sitting). Also, Shlizerman 's model often fails to estimate finer poses such as "hand rising" pose (see the rightmost plot shown with the red arrow). In contrast, our method outputs closer poses to ground truth. We consider that our network model with six convolution layers and two FC layers perceives finer poses compared with the baseline network with a single LSTM and FC layers.

Ablative Analysis. Our method uses two types of acoustic features, , intensity vector and log Mel spectrum, as described in 3.3. This ablation test investigates the effect of these two features for both classification and 3D pose estimation task. To this end, we trained our model in three settings with audio features sampled with 30 fps for the classification, and 10 fps for the pose estimation: (i) full set, (ii) exclude intensity vector, and (iii) exclude log Mel spectrum. As indicated in Table 4, the combination of both two features results in the best estimation accuracy, while intensity vector was of more critical importance regarding performance.

Investigation on FPS for Feature Extraction. As described in 3.3, we extract the acoustic features at a fixed window length (frame rate). A longer window length (i.e., lower frame rate) provides more information per window due to the increase in the number of samples, but at the same time, the temporal resolution per window decreases, which may result in degradation of the estimation accuracy. We empirically investigate the optimal window length. Table 5 shows the results with 10, 20 and 30 fps.

As the table shows, the model trained with the features with 10 fps achieved the highest performance, and the accuracy decreases as fps is increased. However, note that our model trained with 30 fps still outperforms the baseline model trained with 10 fps (see Table 3 for the performance by the baseline), which shows the efficacy of our model.

Evaluation with In Plain Clothes Dataset. To show our approach's possibility to work with real-world data, we further test

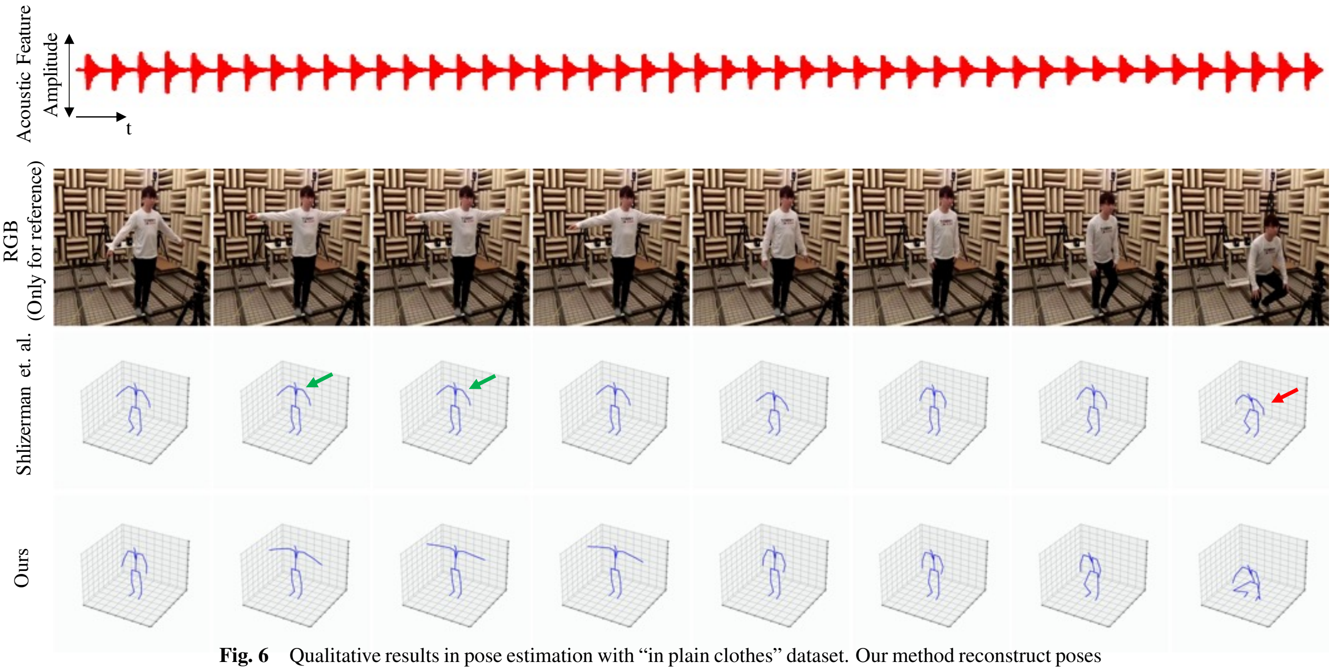


Fig. 6 Qualitative results in pose estimation with "in plain clothes" dataset. Our method reconstruct poses closer to ground truth compared with the baseline method. The results show that our acoustic sensing system and the inferring model are robust to the clothes that the subjects put on.

our method on the "in plain clothes" dataset described in Sec. 4.1 in which the subjects were asked not to wear Mocap suits. As shown in Fig. 6, our approach produces better poses than the baselines in varied cases, including "T-pose" (see the second and third from the left, shown with the green arrows) and "sitting" (see the rightmost plot, annotated with the red arrow). The results show that our acoustic sensing system and the inferring model are robust to the clothes that the subjects put on, which shows the potential of our system to be used in real-world data.

6. Discussion and Limitations

This paper explored the ability to capture human behavior using low-level acoustic signals. While we showed promising results in this paper, there remain some limitations. This section discusses them to give some promising directions to future research.

Frequency Range Used for TSP Signals. For active acoustic sensing, we emitted TSP signals with audible range of signal frequency. However, we are planning to test our system with over 20 kHz, which is inaudible to the human ear and hence the approach can be entirely silent to the user.

Practicality of the Approach. In this paper, as the first step toward tackling the new task, we have conducted experiments in an echoic chamber environment, where the reverberation or other noise can be reduced. However, in order to test the practicality in more general situations, we plan to use data recorded in natural reverberant rooms with more noise and echoes.

Spatial Resolution. Due to the longer wavelength compared with other modalities, , visual spectrum or RF/WiFi, the spatial resolution of the acoustic signal-based approach that we applied is lower than those other methods. Although this characteristic of acoustic signal brings an important advantage of the robustness to occlusions, it results in missing some small motion behaviors of the targets. Our results showed that changes in arms and legs can be captured, but we will verify if more detailed changes (such as hand and head tilt) can be estimated.

Dataset Size. Since we tackled the new task, we captured our own dataset. Compared with datasets captured with conventional cameras [4], [8], our data collection process requires audio signals synchronized with Mocap data, and thus collecting large-scale datasets would be more difficult. We will explore more efficient way to collect larger-scale datasets, as well as to automatically generate synthetic data.

7. Conclusion

This work proposes a framework to infer human behavior, , action categories or 3D poses of humans, given only low-level acoustic signals. Our framework uses audio features that include the direction of arrival of the sound as well as signals that mimic the non-linear human ear perception of sound. As a result, we have shown for the first time that it is possible to take low-level audio signals into a high-level understanding of human behavior aided by the power of the data-driven approach. Though more research is necessary to make the approach practical, we believe that this preliminary work brings to the community a new possibility for acoustic inference of essentially visual information and remarkable potential for higher-level reasoning based on acoustic measurement.

References

[1] Fadel Adib, Chen-Yu Hsu, Hongzi Mao, Dina Katabi, and Frédo Durand. Capturing the human figure through a wall. *ACM Transactions on Graphics (TOG)*, 34(6):1–13, 2015.

[2] Yusuf Aytar, Carl Vondrick, and Antonio Torralba. SoundNet: Learning sound representations from unlabeled video. In *Conference on Neural Information Processing Systems (NeurIPS)*, page 892–900, 2016.

[3] Yin Cao, Turab Iqbal, Qiuqiang Kong, Miguel Galindo, Wenwu Wang, and Mark Plumbley. Two-stage sound event localization and detection using intensity vector and generalized cross-correlation. Technical report, DCASE2019 Challenge, 2019.

[4] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. OpenPose: Realtime multi-person 2D pose estimation using part affinity fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 43(1):172–186, 2021.

[5] Changan Chen, Unnat Jain, Carl Schissler, Sebastia Vicenc Amengual Gari, Ziad Al-Halah, Vamsi Krishna Ithapu, Philip Robinson, and Kristen Grauman. SoundSpaces: Audio-visual navigation in 3D environments. In *European Conference on Computer Vision (ECCV)*, pages 17–36, 2020.

[6] Viviana Crescitelli, Atsutake Kosuge, and Takashi Oshima. An RGB/infra-red camera fusion approach for multi-person pose estimation in low light environments. In *IEEE Sensors Applications Symposium (SAS)*, pages 1–6, 2020.

[7] Travis Deyle, Szabolcs Palinko, Erika Shehan Poole, and Thad Starner. Hambone: A bio-acoustic gesture interface. In *IEEE International Symposium on Wearable Computers (ISWC)*, pages 3–10, 2007.

[8] Hao-Shu Fang, Shuqin Xie, Yu-Wing Tai, and Cewu Lu. RMPE: Regional multi-person pose estimation. In *IEEE International Conference on Computer Vision (ICCV)*, pages 2334–2343, 2017.

[9] Ruohan Gao, Tae-Hyun Oh, Kristen Grauman, and Lorenzo Torresani. Listen to look: Action recognition by previewing audio. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10457 – 10467, 2020.

[10] Shiry Ginosar, Amir Bar, Gefen Kohavi, Caroline Chan, Andrew Owens, and Jitendra Malik. Learning individual styles of conversational gesture. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3497–3506, 2019.

[11] Teng Han, Khalad Hasan, Keisuke Nakamura, Randy Gomez, and Pourang Irani. SoundCraft: Enabling spatial interactions on smartwatches using hand generated acoustics. In *ACM Symposium on User Interface Software and Technology (UIST)*, page 579–591, 2017.

[12] Chris Harrison and Scott E. Hudson. Scratch Input: Creating large, inexpensive, unpowered and mobile finger input surfaces. In *ACM Symposium on User Interface Software and Technology (UIST)*, page 205–208, 2008.

[13] Chris Harrison, Desney Tan, and Dan Morris. Skinput: Appropriating the body as an input surface. In *SIGCHI Conference on Human Factors in Computing Systems (CHI)*, page 453–462, 2010.

[14] Shawn Hershey, Sourish Chaudhuri, Daniel P. W. Ellis, Jort F. Gemmeke, Aren Jansen, R. Channing Moore, Manoj Plakal, Devin Platt, Rif A. Saurous, Bryan Seybold, Malcolm Slaney, Ron J. Weiss, and Kevin W. Wilson. CNN architectures for large-scale audio classification. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 131–135, 2017.

[15] Go Irie, Mirela Ostrek, Haochen Wang, Hirokazu Kameoka, Akisato Kimura, Takahito Kawanishi, and Kunio Kashino. Seeing through sounds: Predicting visual semantic segmentation results from multichannel audio signals. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3961–3964, 2019.

[16] Mariko Isogawa, Ye Yuan, Matthew O'Toole, and Kris M. Kitani. Optical non-line-of-sight physics-based 3D human pose estimation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7013–7022, 2020.

[17] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, 2015.

[18] Yuki Kubo, Yuto Koguchi, Buntarou Shizuki, Shin Takahashi, and Otmar Hilliges. AudioTouch: Minimally invasive sensing of microgestures via active bio-acoustic sensing. In *International Conference on Human-Computer Interaction with Mobile Devices and Services (MobileHCI)*, 2019.

[19] Jing Li, Di Kang, Wenjie Pei, Xuefei Zhe, Ying Zhang, Zhenyu He, and Linchao Bao. Audio2Gestures: Generating diverse gestures from speech audio with conditional variational autoencoders. In *IEEE International Conference on Computer Vision (ICCV)*, pages 11293–11302, 2021.

[20] Tianhong Li, Lijie Fan, Mingmin Zhao, Yingcheng Liu, and Dina Katabi. Making the invisible visible: Action recognition through walls and occlusions. In *IEEE International Conference on Computer Vision (ICCV)*, pages 872–881, 2019.

[21] Brian McFee, Colin Raffel, Dawen Liang, Daniel PW Ellis, Matt McVicar, Eric Battenberg, and Oriol Nieto. librosa: Audio and music signal analysis in Python. In *Proc. 14th Python in Science Conference*, 2015.

[22] Adiyan Mujibiya, Xiang Cao, Desney S. Tan, Dan Morris, Shwetak N. Patel, and Jun Rekimoto. The sound of touch: On-body touch and gesture sensing based on transdermal ultrasound propagation. In *ACM International Conference on Interactive Tabletops and Surfaces (ITS)*, page 189–198, 2013.

[23] Senthil Purushwalkam, Sebastian Vicenc Amengual Gari, Vamsi Krishna Ithapu, Carl Schissler, Philip Robinson, Abhinav Gupta, and Kristen Grauman. Audio-visual floorplan reconstruction. In *IEEE International Conference on Computer Vision (ICCV)*, pages 1183–1192, 2020.

[24] Fengmin Shi, Jie Guo, Haonan Zhang, Shan Yang, Xiying Wang, and Yanwen Guo. GLAVNet: Global-local audio-visual cues for fine-grained material recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14433–14442, 2021.

[25] Eli Shlizerman, Lucio M Dery, Hayden Schoen, and Ira Kemelmacher-Shlizerman. Audio to body dynamics. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7574–7583, 2017.

[26] Fei Wang, Stanislav Panev, Ziyi Dai, Jinsong Han, and Dong Huang. Can WiFi estimate person pose? *ArXiv*, abs/1904.00277, 2019.

[27] Tomohiro Yokota and Tomoko Hashida. Hand gesture and on-body touch recognition by active acoustic sensing throughout the human body. In *Symposium on User Interface Software and Technology (UIST)*, page 113–115, 2016.

[28] Mingmin Zhao, Tianhong Li, Mohammad Abu Alsheikh, Yonglong Tian, Hang Zhao, Antonio Torralba, and Dina Katabi. Through-wall human pose estimation using radio signals. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, pages 7356–7365, 2018.