Listening Human Behavior: 3D Human Pose Estimation with Acoustic Signals

Yuto Shibata
Keio University
yuto071508@keio.jp

Yutaka Kawashima
Keio University
ykawashima@aoki-medialab.jp

Mariko Isogawa
Keio University, JST Presto
mariko.isogawa@keio.jp

Go Irie
Tokyo University of Science
goirie@ieee.org

Akisato Kimura
NTT Corporation
akisato@ieee.org

Yoshimitsu Aoki
Keio University
aoki@elec.keio.ac.jp

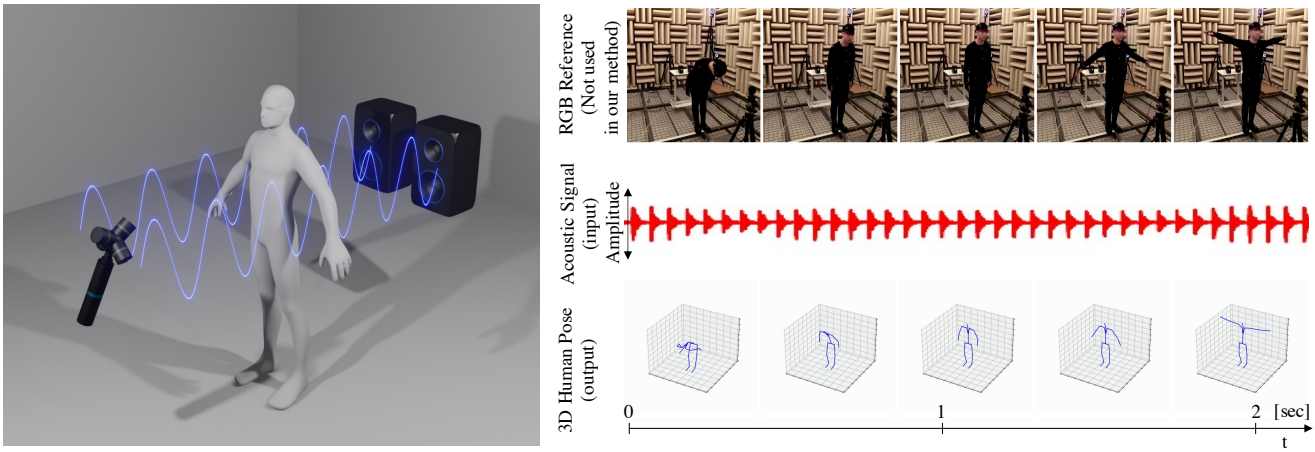

Figure 1. We propose 3D human pose estimation given only low-level acoustic signals with a single pair of microphones and loudspeakers. Given an audio feature frame (right-middle), our method estimates 3D human pose sequences (right-bottom).

Abstract

Given only acoustic signals without any high-level information, such as voices or sounds of scenes/actions, how much can we infer about the behavior of humans? Unlike existing methods, which suffer from privacy issues because they use signals that include human speech or the sounds of specific actions, we explore how low-level acoustic signals can provide enough clues to estimate 3D human poses by active acoustic sensing with a single pair of microphones and loudspeakers (see Fig. 1). This is a challenging task since sound is much more diffractive than other signals and therefore covers up the shape of objects in a scene. Accordingly, we introduce a framework that encodes multichannel audio features into 3D human poses. Aiming to capture subtle sound changes to reveal detailed pose information, we explicitly extract phase features from the acoustic signals together with typical spectrum features and feed them into our human pose estimation network. Also, we show that reflected or diffracted sounds are easily influenced by subjects' physique differences e.g., height and muscularity, which deteriorates prediction accuracy. We reduce these gaps by using a subject discriminator to improve accuracy. Our experiments suggest that with the use of only low-dimensional acoustic information, our method outperforms baseline methods. The datasets and codes used in this project will be publicly available.

1. Introduction

The ability to capture human behavior, such as 3D poses, has many potential applications. Over the last decade, many different technologies have been proposed to infer human poses, including conventional cameras [4, 8], transient light [16], radio frequency (RF) or WiFi measurements [22, 33]. However, the optical signals are easily occluded and restricted by poor lighting conditions, 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, 22, 30, 33]	RF/WiFi	Metal, water	Low	Non-invasive
Audio to joint [10, 21, 28]	Audio	Soundproof room	High (speech required)	Non-invasive
Audio to hand micro gesture [20]	Audio	Soundproof room	Low	Invasive
Ours	Audio	Soundproof room	Low	Non-invasive

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

a dark room or a night road. RF/WiFi signals are also occluded by water or metal. In addition, the use of wireless signals is often limited, as electronic devices that transmit 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 waves, which are outside of our audible range, we are not even aware of them. Moreover, since 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 in cross-modal analyse with visual information, including scene geometry estimation [5, 26], action recognition [9], visual semantic segmentation [15], and even object understanding [27]. Another line of studies uses acoustic signals for sensing humans, such as active hand gesture monitoring [20]. Papers that are more relevant to ours are those that infer human joints by converting human speech or music to gestures [10, 21, 28]. These methods use human speech as a clue for recovering human gestures/motions. However, these methods utilize semantics of sounds, for instance, human voice, music, speech, or sounds of specific actions, which raise privacy issues. In this paper, we define such signals with the semantics of sounds as "high-level signals" and signals that do not include any of the semantics of sounds as "low-level signals." So far, no methods have been proposed to capture whole human 3D poses given only low-level acoustic signals.

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

To answer these questions, this paper examines a new task, *3D human pose estimation from only low-level acoustic signals*. This is a challenging task. The wavelengths of acoustic signals are 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 (Fig. 1), as opposed to previous methods, which 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 in a specific single position and has far fewer geometry clues to map 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 multichannel audio features as its inputs and directly output the predicted 3D body part joint locations. If humans occlude acoustic signals emitted from loudspeakers, this subtle "shifts" in arrival time of the incoming acoustic signals will occur. Our network model captures these small shifts by explicitly integrating phase features that represent the time difference of arrival (TDOA) and utilizing them to infer human behavior. Additionally, we discover that sounds reflected or diffracted on subjects' bodies tend to be affected by their physique differences, such as height and muscularity. Our proposed dataset contains sound data of both men and women, and the physical features vary among subjects. This difference causes our model to over-fit on each subject's physical characteristics and prevents it from generalizing well to unseen subjects. Therefore, we apply adversarial learning to this task using the subject discriminator's prediction uncertainty and create subject-invariant features.

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. These data were recorded in both (i) an anechoic room with little effect of reverberation and (ii) a classroom with a lot of noise.

To summarize, our contributions are as follows: (1) We are the first to tackle a new task: 3D human pose estimation given only low-level audio signals; (2) We describe a network architecture that directly maps acoustic features to 3D human poses; (3) We increase prediction accuracy by using

adversarial learning and creating human-physique-invariant features; (4) Since there are no previous methods to carry out this task, we describe how to create new datasets to train our network model; and (5) We conduct 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. This section introduces them and describes other relevant approaches in detail.

Human Pose Estimation with Different Modalities. Estimating human poses has long been a research topic in the computer vision community [4, 8]. Although a majority of existing work leverages the fact that the human body is visible to cameras, this line of research includes a wide variety of solutions in terms of hardware systems and reconstruction algorithms that operate in different parts of the spectrum. Besides the visible spectrum (380–740 nm) [16] or near-IR (740–1500 nm) light [6], WiFi and RF (centimeter scale) [22, 33] or even sound waves (meter scale) [20] 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 (*e.g.*, a dark room, night road) and occluded by other objects (*e.g.*, buildings, etc.). RF or WiFi signals enable through-the-wall pose estimation [1, 22, 33] since longer electromagnetic waves tend to pass through objects; however, these signal spectra are also occluded by some materials (*i.e.*, metal, water), and their use is often limited to being used. For example, electronic devices that transmit signals must be turned off during flights as well as in hospital rooms with sensitive electronic systems.

Acoustic signals have the potential to overcome the issues raised above. They are less affected by occlusion than other signals due to their longer wavelength. Additionally, acoustic signals are rarely prohibited, unlike wireless signals. However, this also presents a number of fundamental limitations when estimating human poses. For example, the spatial and angular resolution must be limited, making it hard to distinguish small differences in poses. We will address these limitations and explore the potential of active acoustic sensing for 3D human pose estimation.

Acoustic Sensing for Capturing Human Behavior. In this paper, we leverage active acoustic sensing using a single pair of ambisonics microphones and loudspeakers. Hence, our work is closely related to studies that leveraged acoustic sensing for capturing human behavior. Passive acoustic sensing has been used for gesture recognition [7, 11, 12], on-body sensing [13], activity [9], or even body joints estimation [10, 21, 28]. However, these methods require their users to put on wearable devices [7, 11–13], and use high-level acoustic information, such as human speech [10, 21, 28] or daily activity sound [9]. These signals contain a great deal of information, which raises privacy issues. Moreover, these approaches basically reconstruct "gestures" and are thus not designed to estimate more "subtle" differences in poses. This is because passively obtained acoustic signals often lack enough information to recover finer human poses. Active acoustic sensing has also been researched to provide a more detailed level of gesture recognition [20, 25, 32]. However, these methods also require the placement of devices on a part of a human body to capture its movement or position. Inspired by these previous works, we capture 3D human poses in a non-invasive manner, given low-level acoustic signals only.

3. Methodology

Given only a sequence $\mathbf{s} = [s_1, s_2, ..., s_T]$ of low-level acoustic signals, our goal is to infer the 3D human pose sequence $\mathbf{p} = [p_1, p_2, ..., p_T]$, where p_t represents the 3D joint position of frame t. Here, T indicates the number of samples. An overview of our method is provided in Fig. 2. Our proposed framework consists of an acoustic feature extraction module that encodes raw acoustic signals into a sequence of acoustic feature vectors $\mathbf{a} = [a_1, a_2, ..., a_T]$ and 3D human pose estimation network f. The following subsections describe the active acoustic sensing (Sec. 3.1) and acoustic features that are fed into f (Sec. 3.2). Then, Sec. 3.3 explains our 3D human pose estimation network.

3.1. Active Acoustic Sensing

Suppose we have a known sound source and a microphone. The sound emitted from the source bounces off objects in 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 by the source when 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 a successful existing active acoustic sensing technique [20], we transmit a modulated acoustic signal and pre-process the received signal to emulate RIR. This is a similar approach to the "chirp signal" generally applied to FMCW radar, which transmits linear sweep frequency-modulated signals. We specifically use a time-stretched pulse (TSP) as our sound $s'(t)$, which is a kind of swept sine wave designed for RIR measurement.

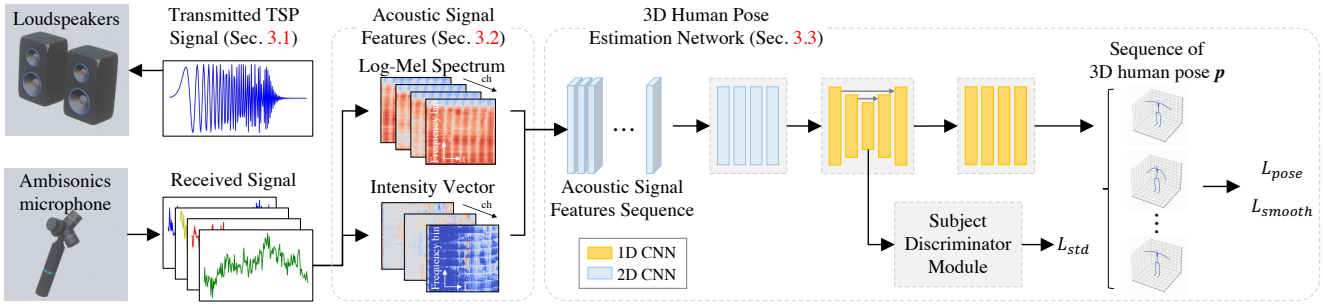

Figure 2. The overview of our framework for acoustic signal-based 3D human pose estimation.

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

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 the 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 a TSP signal with a sampling rate of 48 kHz, and its entire waveform length N was set to 4096.

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 microphone polar pattern, pointing in a specific direction.

3.2. Acoustic 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 $I^{\text{intensity}}$ that has $b \times 3$ dimensions, 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 [31] 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 in validation and testing. 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, *i.e.*, 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 perceiving the scene geometry. In order to concatenate the intensity vector and the log Mel spectrum that we describe later, following [3], we compute the intensity vector in the short-time Fourier transform (STFT) domain and the Mel space as follows:

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

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

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 the $L1$ norm. We then standardize $\hat{I}(k,t)$ to extract final intensity vector $I^{\text{intensity}}$ that we input into the network.

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

$$I^{\text{mel}}(k,t) = H_{\text{mel}}(k,f) \cdot \mathcal{F}(s(f,t)), \quad (4)$$

where k is the index of the Mel bins, H_{mel} represents the Mel-bank filter, and $\mathcal{F}$ is the Fourier transform operation. We then convert it to log scale and standardize it to obtain the feature I^{logmel}, which is also fed into the network.

3.3. 3D Human Pose Estimation Network

Next, we introduce our 3D human pose estimation network f that outputs human pose $p = f(a)$ given acoustic feature vector a. The network f includes the *subject discriminator module* that reduces subjects' physique differences, aiming to increase the generalization ability for unseen subjects.

Pose Estimator Network. Following a previous work using acoustic signals, we use a similar architecture to [10], which consists of a sequence of acoustic features as input and time-wise U-Net for temporal consistency-aware pose estimation. Unlike this previous work, we leverage the intensity vector so that the network can infer finer human poses based on the direction of sound arrival. In addition, compared with [10], our datasets contain more dynamic movement. Therefore, we changed the time scale (time window for Fourier transform and sequence length). More specifically, the network takes 12 frames of the acoustic feature sequence as input, which has a size of $7 \times 128 \times 12$. It is then fed into a 2D CNN with four blocks to generate temporal consistency-aware acoustic feature ϕ with a size of $4096 \times 12 \times 1$. Then, each row of ϕ is fed into a U-Net shaped network with five 1D CNN layers to generate pose feature ρ, and finally, after going through four 1D CNN layers, the network output pose p with a size of 12×63, which consists of 3D $\times 21$ joints for each 12 frames.

With the variable θ that contains all trainable parameters, the training objective uses Mean Squared Error (MSE) loss

$$\mathcal{L}_{pose}(\theta) = \frac{1}{T}\Sigma_i^T(\hat{p}_i - p_i)^2. \quad (5)$$

Here, $\mathcal{L}$ denotes the loss function, and $\hat{p}$ represents the ground-truth position of the pose. In addition to $\mathcal{L}_{pose}$, following [18], we also use smooth loss to make our prediction smoother:

$$\mathcal{L}_{smooth}(\theta) = \frac{1}{T-1}\Sigma_{i=2}^T||(\hat{p}_i - \hat{p}_{i-1}) - (p_i - p_{i-1})||. \quad (6)$$

Subject Discriminator Module. Regarding acoustic signal-based human behavior estimation, it is reported that the estimation performance is highly dependent on *domain* [17, 18]. Here, *domain* expresses a pair of human bodies and a recording environment. With the presence of such a difference of domain (hereafter, domain gap), prediction accuracy is easily affected by subjects' physique differences, and hence our model has a reduced generalization ability to predict the joint positions of unseen subjects. Therefore, we introduce "subject discriminator," an adversarial learning-based module, to remove subject-specific features and leverage only human-invariant information.

Our discriminator has a single fully connected layer. The input of the discriminator is the output of i-th hidden layer of the U-Net-shaped network inside f. As shown in Fig. 2, we set this i to the number of the most pooled layer so that we can leverage information about the entire sequence. Given the hidden layer output, the discriminator outputs subject distribution S. Based on the S, our discriminator is trained using cross-entropy loss so it can distinguish each subject. Unlike previous work [17], we leverage the uncertainty of the subject discriminator's prediction, namely, the standard deviation of its output. We define the standard deviation (STD) loss as follows:

$$L_{std} = \frac{1}{T'}\Sigma_{n=1}^{T'}\mathrm{STD}(S_n), \quad (7)$$

where S_n and T' denote the n-th sequence's prediction regarding subject distribution and the number of total sequences respectively. The function $\mathrm{STD}(\cdot)$ computes the standard deviation of the elements. Based on Eqs.(5), (6), and (7), our final loss function is represented as follows with the weight parameters w_α, w_β, and w_γ:

$$L = w_\alpha L_{pose} + w_\beta L_{smooth} + w_\gamma L_{std}. \quad (8)$$

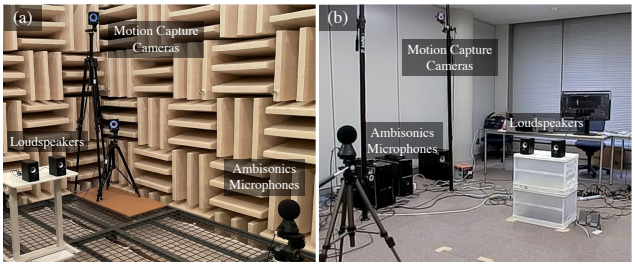

Figure 3. The setup of our experiments in (a) an anechoic chamber, and (b) a classroom.

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. 3, we used a pair of ambisonics microphone (Zoom H3-VR) and loudspeakers (Sanwa Supply MM-SPU9BK). The acoustic signals were captured in both (a) an anechoic chamber environment, where the reverberation or other noise can be reduced, and (b) an ordinal classroom in a university containing noises and reverberation. Both datasets are one hour in length (equal to 3.6K frames at 10 fps). They consist of eight subjects[1] who were asked to wear the Mocap suit and stand between a microphone and loudspeakers while the subjects performed the actions.

The subjects performed various complex poses: walking, sitting, bending forward, raising both hands, and transitioning between all of these motions. For pose ground-truth annotation, we used the skeleton of 21 joints, including the

[1] All subjects agreed that their Mocap data and captured acoustic signals could be used in this research.

Table 2. Quantitative experimental results in the anachoic chamber environment (left half) and the classroom environment (right half).

Method	Anechoic Chamber Environment						Classroom Environment					
	Single Subject			Cross Subject			Single Subject			Cross Subject		
	RMSE (↓)	MAE (↓)	PCKh @0.5 (↑)	RMSE (↓)	MAE (↓)	PCKh @0.5 (↑)	RMSE (↓)	MAE (↓)	PCKh @0.5 (↑)	RMSE (↓)	MAE (↓)	PCKh @0.5 (↑)
Ginosar *et al*. [10]	0.44	0.23	0.90	0.83	0.51	0.60	0.58	0.30	0.84	0.95	0.56	**0.68**
Jiang *et al*. [18]	0.90	0.44	0.73	0.96	0.55	0.62	0.58	0.34	0.73	1.02	0.63	0.49
Ours (Method's best)	**0.42**	**0.22**	**0.90**	**0.73**	**0.45**	**0.72**	**0.54**	**0.28**	**0.85**	**0.93**	**0.55**	0.67

head, neck, shoulders, arms, forearms, hands, hips, legs, feet, toes, pelvis, and spine. These datasets include subjects with various physical scales. Therefore, we applied additional translation so that the position of the hips sits at the origin of our coordinates, followed by normalization using the length between the spine and hips. These datasets were used for both training and testing.

In Plain Clothes Dataset. To further showcase our method's applicability, we also tested our method on the acoustic signal captured with subjects not wearing 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 those wearing Mocap suits. Since these datasets do not include ground-truth poses, these were only used 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 sounds, such as human speech or music. Therefore, we compared our network model against the following similar approaches to ours; (i) Ginosar *et al*. [10] that uses audio signals like ours but includes high-level signal, *i.e*., human speech, and (ii) Jiang *et al*. [18] one of the state-of-the-art methods for capturing 3D human poses from actively captured low-level signals, *i.e*., WiFi signals. Both networks were trained with our datasets for fair architecture comparisons. To this end, we modified the input and the last layer of the baseline network so that they can use our acoustic signals as input, and outputs 3D poses. For Jiang *et al*.'s method, although in their original work they predicted angles between joints thus getting positions using known body shapes, we modify the method to directly predict the positions of joints. This is because as mentioned in 4.1, we apply positional normalization with the root position to the samples, making the skeleton structure have a certain level of correctness without using any known body shape.

4.3. Evaluation Metrics

We used 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. PCK measures the percentage of the predicted joint locations that are within a specific range from the ground truth. Specifically, this paper applied the PCKh@0.5 score, which uses a threshold of 50% for the head–neck bone link.

4.4. Implementation Details

Audio Signal Features. We used librosa [24] as an audio signal processing library to extract the log Mel spectrum. We sampled acoustic frames to extract features at 20 fps.

Networks and Training. In all experiments, we used Adam [19] to optimize our network. We set the learning rate to 0.003 and 0.001 with and without the subject discriminator module respectively. The network function typically converges after 100 and 30 epochs in single-subject and cross-subject settings, respectively. As for weight parameters, we set $w_\alpha = w_\beta = 1.0$. $w_\gamma = 0.5$ and $w_\gamma = 1.0$ were set for an anechoic chamber and classroom environment, respectively. We set $w_\gamma = 0.5$ for our model without the intensity vector, which will be described in an ablative analysis.

5. Experiments and Results

We conducted five different experiments to investigate our method's efficacy: (1) a comparison against existing pose estimation baselines; (2) an ablation study to show the importance of the intensity vector and subject discriminator, which are our main technical contributions; (3) a comparison between our subject discriminator loss and the existing discriminator loss method; and (4) an investigation of the trade-off between the length of the time window and estimation accuracy. While these four tests used a motion capture suit dataset that includes ground truth, we also conducted (5) qualitative analyses with the "in plain clothes" dataset.

Comparison against baseline methods. We compared our

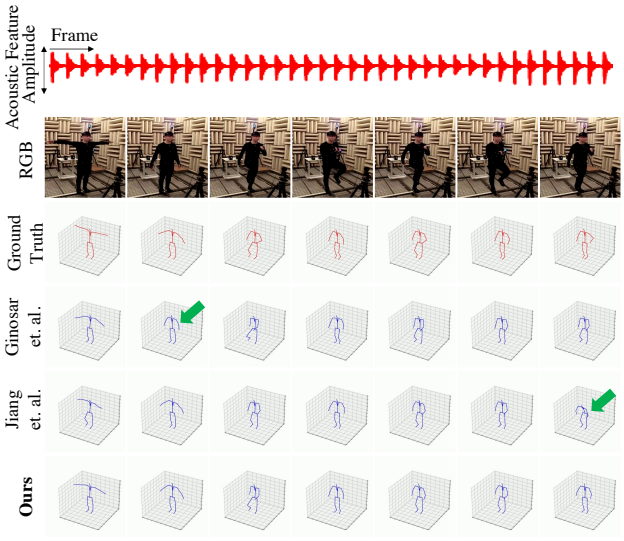

Figure 4. Qualitative results with the motion capture suit dataset. While the baseline method failed to reconstruct finer poses, our method output closer poses to the ground truth.

method against the baselines with the two different datasets to show our method's practicality with sound noises and reverberation. Both experiments were conducted in two different settings: (i) a single-subject setting, using the same subject for both training and testing; and (ii) a cross-subject setting, using different subjects for training and testing. For the single-subject setting, we trained our model for each subject. We used an 80:20 train-test data split in a time-series manner. For the cross-subject setting, we trained models on seven subjects and tested them on the other subject. Please note that the subject discriminator module was only introduced for the cross-subject setting. This is because the single-subject setting does not require any physical differences between the subjects to be removed.

Table 2 (left) shows the quantitative results. Here, we compare our method's best model (i.e., our method with intensity vector, except for cross-subject/classroom. Subject discriminator is also introduced in a cross-subject setting) with baselines. The reason why the intensity vector is removed in a cross-subject/classroom setting is explained in the next experiment for ablative analysis. As shown in the table, our method outperformed other baselines in all the settings except PCK@0.5 in the classroom/cross-subject setting. Fig. 4 shows the qualitative results of the cross-subject setting[2]. As the green arrows show, Ginosar *et al.*'s model was not able to reproduce finer human poses like "hand rising." Moreover, Jiang *et al.*'s model often predicted the average shape. In contrast, our method outputs poses that were closer to ground truth. Hence, our network model with the intensity vector and subject discriminator

[2]Single-subject results are given in the supplementary materials.

Table 3. Ablation study.

Method	Anechoic Chamber			Classroom		
	RMSE (↓)	MAE (↓)	PCKh @0.5 (↑)	RMSE (↓)	MAE (↓)	PCKh @0.5 (↑)
Ours	**0.73**	**0.45**	**0.72**	0.97	0.57	0.64
Ours w/o SM	0.79	0.48	0.68	0.98	0.57	0.67
Ours w/o IV	0.77	0.47	0.68	**0.93**	**0.55**	**0.67**

Table 4. Comparison between STD and ordinal discriminator loss.

Method	RMSE (↓)	MAE (↓)	PCKh@0.5 (↑)
Ours (L_d [17])	0.78	0.47	0.68
Ours (L_{std})	**0.73**	**0.45**	**0.72**

module perceives finer poses of the unseen subject compared with the baseline networks.

Ablative Analysis. This ablation test investigated the effect of (a) using an intensity vector and (b) introducing a subjective discriminator module, which are our main technical contributions. As shown in Table 3, in the anechoic chamber environment, our model reported the best score compared with both our model without the subject discriminator module (denoted as Ours w/o SM) and our model without the intensity vector (denoted as Ours w/o IV), indicating that both essences contributed to improving the estimation accuracy. In the classroom environment, Ours w/o IV outperformed the other implementations. We believe that this is because, unlike an anechoic chamber, a normal classroom contains various noises and reverberations, which makes the intensity vector too noisy to capture the arrival direction of the sound precisely.

Effect of standard deviation loss on the subject discriminator module. The previous paragraph explained that our subjective discriminator module contributes to total scores. However, one might wonder if our standard deviation (STD) loss L_{std} is more effective than other discriminators. Here, we compare the performance of the model with our STD loss compared against the discriminator used in the previous work [17] in a cross-subject setting in the anechoic room environment. The previous work [17] utilizes environment-invariant features and incorporates the negative cross-entropy loss of the discriminator into the final objective. We denote this loss function with their discriminator as L_d, and we used it in place of L_{std} for comparison. For L_d, we set $w_\gamma = 0.1$.

The results are shown in Table 4. Based on these results, our method with the subjective discriminator that uses L_{std} (denoted as Ours(L_{std})) outperformed the method with L_d (denoted as Ours(L_d)), which indicates that our loss L_{std} works more effectively. Also, Fig. 5 illustrates that our subjective discriminator module removed sample feature differences among varied subjects. Samples from three different

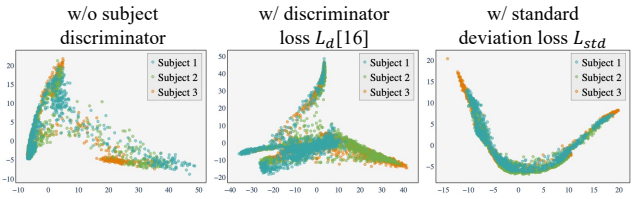

Figure 5. The intermediate outputs of three subjects.

subjects, including 2 males and 1 female, were dimensionally reduced into 2D. As shown in the figure, the samples without any subject discriminator (left) and those with L_d (middle) exhibited large differences among subjects, which were successfully removed in samples with L_{std} (right).

Investigation of FPS for Feature Extraction. As described in 3.2, we extracted 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 investigated the optimal window length. Table 5 shows the results at 20, 30, and 40 fps in an anechoic room and in the cross-subject setting. As the table shows, the model trained with the features at 20 fps achieved the highest performance, and PCKh@0.5 decreased as fps increased. For RMSE and MAE, there was no significant negative trend as fps increased. However, note that our model trained at 40 fps still outperformed the baseline models trained at 20 fps (see Table 2 for the baseline performamnce), which shows the efficacy of our model.

Evaluation with In Plain Clothes Dataset. To show the applicability of our approach to real-world data, we further tested our method on the "in plain clothes" dataset described in Sec. 4.1, in which the subjects were asked not to wear Mocap suits in an anechoic chamber environment. As shown in Fig. 6, in the cross-subject setting, our method produced better poses than the baselines in various cases, including "T-pose" (see the first line with green arrow), and successfully avoided average prediction (see the second line with a green arrow).

Additionally, we employed the Strided Transformer [23] as an image-based pose prediction model to obtain pseudo labels for our plain clothes dataset within a classroom environment. The results are displayed in Table 6, illustrating practicality in plain clothes and noisy settings. The Strided Transformer was trained using Human 3.6M [14] and HumanEva-I [29] datasets, which have slightly different definitions regarding the positions of each joint compared to our Mocap datasets. Furthermore, this image-based model introduces some errors in the z-dimension. Consequently, the results in 6 contain some unavoidable errors; thus, please consider this quantitative data as a reference only.

Table 5. Investigation on FPS for feature extraction.

Method	RMSE (↓)	MAE (↓)	PCKh@0.5 (↑)
Ours (FPS = 20)	**0.73**	**0.45**	**0.72**
Ours (FPS = 30)	0.81	0.50	0.68
Ours (FPS = 40)	0.76	0.46	0.66

Table 6. Quantitive result on our plain clothes dataset

Method	RMSE (↓)	MAE (↓)
Ours best	1.59	1.09

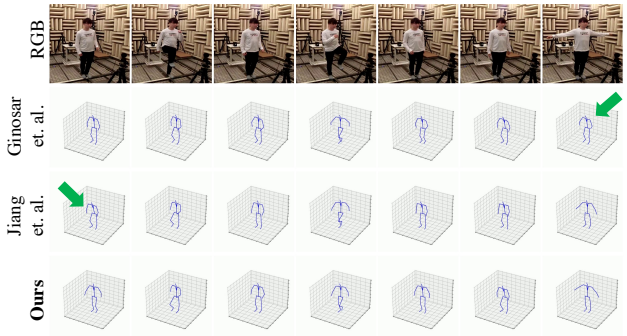

Figure 6. Qualitative pose estimation results with the "in plain clothes" dataset.

6. Limitations

This paper explored the ability to capture human behavior using low-level acoustic signals. While we showed promising results in this paper, there are some limitations. For sensing, we emitted TSP signals within the audible range of signal frequency. However, we plan to test our system with signals over 20 kHz, which are inaudible to the human ear, and hence the approach can be entirely silent for the user.

7. Conclusion

This work proposes a framework to infer 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's perception of sound. We show for the first time that it is possible to use low-level audio signals to obtain a high-level understanding of human behavior aided by the power of the data-driven approach. Further, we found that acoustic features are dependent on the subject's physique and proposed subject discriminator module to extract subject-invariant features. Although more research is necessary to make this approach practical, we believe that this preliminary work offers a new possibility for acoustic inference of essentially visual information and a remarkable potential for higher-level reasoning based on acoustic measurement.

Acknowledgements. This work was partially supported by JST Presto JPMJPR22C1 and Keio University Academic Development Funds.

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, 3

[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. 4

[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

[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. 1, 2, 3

[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. 2

[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. 3

[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. 3

[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. 1, 2, 3

[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. 2, 3

[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. 2, 3, 5, 6

[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. 3

[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. 3

[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. 3

[14] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3. 6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE transactions on pattern analysis and machine intelligence*, 36(7):1325–1339, 2013. 8

[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. 2

[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. 1, 3

[17] Wenjun Jiang, Chenglin Miao, Fenglong Ma, Shuochao Yao, Yaqing Wang, Ye Yuan, Hongfei Xue, Chen Song, Xin Ma, Dimitrios Koutsonikolas, Wenyao Xu, and Lu Su. Towards environment independent device free human activity recognition. In *International Conference on Mobile Computing and Networking (MobiCom)*, page 289–304, 2018. 5, 7

[18] Wenjun Jiang, Hongfei Xue, Chenglin Miao, Shiyang Wang, Sen Lin, Chong Tian, Srinivasan Murali, Haochen Hu, Zhi Sun, and Lu Su. Towards 3d human pose construction using wifi. In *International Conference on Mobile Computing and Networking (MobiCom)*, pages 1–14, 2020. 5, 6

[19] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, 2015. 6

[20] Yuki Kubo, Yuto Koguchi, Buntarou Shizuki, Shin Takahashi, and Otmar Hilliges. AudioTouch: Minimally invasive sensing of micro-gestures via active bio-acoustic sensing. In *International Conference on Human-Computer Interaction with Mobile Devices and Services (MobileHCI)*, 2019. 2, 3

[21] 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. 2, 3

[22] 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. 1, 2, 3

[23] Wenhao Li, Hong Liu, Runwei Ding, Mengyuan Liu, Pichao Wang, and Wenming Yang. Exploiting temporal contexts with strided transformer for 3d human pose estimation. *IEEE Transactions on Multimedia*, 2022. 8

[24] 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. 6

[25] 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. 3

[26] 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. 2

[27] 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. 2

[28] 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. 2, 3

[29] Leonid Sigal, Alexandru O Balan, and Michael J Black. Humaneva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion. *International journal of computer vision*, 87(1-2):4, 2010. 8

[30] Fei Wang, Stanislav Panev, Ziyi Dai, Jinsong Han, and Dong Huang. Can WiFi estimate person pose? *arXiv preprint*, 2019. 2

[31] Masahiro Yasuda, Yuma Koizumi, Shoichiro Saito, Hisashi Uematsu, and Keisuke Imoto. Sound event localization based on sound intensity vector refined by dnn-based denoising and source separation. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 651–655, 2020. 4

[32] 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. 3

[33] 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. 1, 2, 3