FOOTSTEP DETECTION AND CLASSIFICATION USING DISTRIBUTED MICROPHONES

Kazuhiro Nakadai

Honda Research Institute Japan Co., Ltd./
Grad. Sch. of Creative Sci. and Eng.,
Waseda University
8-1 Honcho, Wako-shi, Saitama, Japan
nakadai@jp.honda-ri.com

Yuta Fujii, Shigeki Sugano

Grad. Sch. of Creative Sci. and Eng.,
Waseda University
3-4-1 Okubo, Shinjuku, Tokyo, Japan
sugano@paradise.mech.waseda.ac.jp

ABSTRACT

This paper addresses footstep detection and classification with multiple microphones distributed on the floor. We propose to introduce geometrical features such as position and velocity of a sound source for classification which is estimated by amplitude-based localization. It does not require precise inter-microphone time synchronization unlike a conventional microphone array technique. To classify various types of sound events, we introduce four types of features, i.e., time-domain, spectral and Cepstral features in addition to the geometrical features. We constructed a prototype system for footstep detection and classification based on the proposed ideas with eight microphones aligned in a 2-by-4 grid manner. Preliminary classification experiments showed that classification accuracy for four types of sound sources such as a walking footstep, running footstep, handclap, and utterance maintains over 70% even when the signal-to-noise ratio is low, like 0 dB. We also confirmed two advantages with the proposed footstep detection and classification. One is that the proposed features can be applied to classification of other sound sources besides footsteps. The other is that the use of a multi-channel approach further improves noise-robustness by selecting the best microphone among the microphones, and providing geometrical information on a sound source.

1. INTRODUCTION

Computational Auditory Scene Analysis (CASA) [1] which aims at understanding a general sound, is essential to a system which interacts with people, since it makes the system understand the surrounding environment robustly. This paper addresses footstep detection and classification as the first step to CASA. A footstep is one of the most important sound sources in a real environment, because it includes various types of information unique to a person, e.g., motion by footstep tracking, human ID by footstep classification, internal states like emotion through walking style recognition, and so on. Thus, many studies on footstep detection and classification have been reported. They are classified into two approaches; single-channel and multi-channel.

A single-channel approach is conventional, which uses a single microphone [2, 3, 4, 5, 6]. They basically focused on features for footstep detection, e.g., spectral features [2], Cepstral features which are common in *Automatic Speech Recognition (ASR)* [4, 6], rhythm as a temporal feature [3], integration with another modality such as seismic information [5]. Although these studies showed high performance and some were applied to a surveillance system, they still have defects as follows:

- The features are only for detecting footsteps.
- The footstep detection range is limited.
- No geometrical information is used.

The first problem is essential in terms of CASA, because there are various types of sound sources in a real environment. Some would have similar characteristics to footsteps, and might be misrecognized as footsteps when features are only for footstep detection. The other two problems are due to the use of a single microphone, and thus multi-channel approaches have been studied [7, 8]. They use conventional microphone array techniques such as *MUltiple SIgnal Classification (MUSIC)* [9], and showed that geometrical information is useful and the detection range becomes wider. However, microphone array techniques such as MUSIC and beamforming basically requires precise time synchronization within submillisecond between microphones, which means that a special multi-channel A/D device is necessary. For time synchronization based on wireless connection and a network, *Network Time Protocol (NTP)* is often used for time synchronization, but it sometimes has a time delay in seconds. Although *Precise Time Protocol (PTP)* is also proposed for more accurate synchronization, it basically requires special hardware. This means that each microphone has to have a direct wired connection with the special multi-channel A/D device. It results in the inflexible system with multiple wired connections. In addition, the length of each wire tends to be long, and thus, noise is easy to be contained with such long wiring. Furthermore, microphone array techniques require expensive computational power. For instance, MUSIC uses eigenvalue decomposition to localize sound, and simple beamforming still uses matrix multiplication. Therefore, it is essential to be free from precise time synchronization even when multiple microphones are adopted.

This paper proposes to use amplitude-based localization with multiple microphones distributed on the floor, which demands only rough time synchronization of about 100 ms. Also, we discuss more general features which are applicable to detect and identify other sound sources by introducing time-domain, spectral, Cepstral and geometrical features.

2. SYSTEM ARCHITECTURE

Fig.1 illustrates the system architecture for footstep detection, which consists of microphones, sound acquisition, sound event detection, feature extraction, and classification. Eight microphones were aligned in a 2×4 grid manner and the distance between the microphones was 1.2 m, which was decided so that at least two microphones could capture a footstep sound with sufficient power. Sound capturing

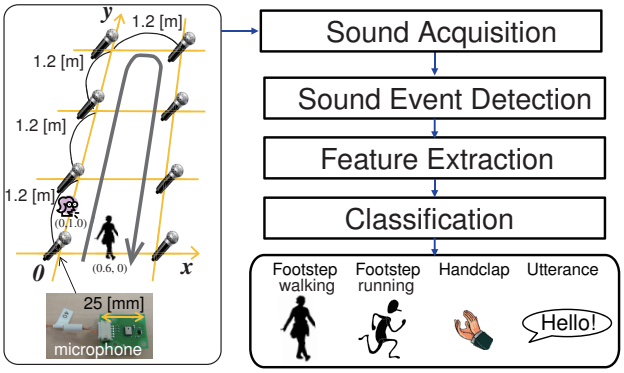


Fig. 1. System Architecture for Footstep Detection

records sounds with the microphones. Sound event detection extracts a sound event one by one in an offline manner. Feature extraction computes a feature vector from each sound event. Classification identifies the sound event as either a walking footstep, running footstep, handclap, or utterance.

The reason of selecting these four sound events are as follows: All of these sound events are related to human behaviors, which means that these are often observed in a real environment, and should be distinguished. Both walking and running are footsteps, but they should have different features such as sound source velocity. Also, handclap has similar characteristics to footsteps. Thus, it may be difficult to distinguish these three similar events properly. In case of the utterance which is a continuous event, it will be interesting to investigate whether our proposed features are effective for utterance-like sound events.

2.1. Sound Event Detection

Sound event detection is based on peak detection in the time domain. To improve the robustness of peak detection, we take a coarse-to-fine approach. First, a sequence of peaks, $\boldsymbol{P} = \{\boldsymbol{p}(i)|1 \leq i \leq N\}$, is detected from a channel-integrated rectified signal,

$$s(n) = \sum_{m=1}^{M} |s_m(n)|, \quad (1)$$

$$\boldsymbol{P} = \boldsymbol{findpeaks}\left(s(n), mindis, bgnlevel\right), \quad (2)$$

where mindis and bgnlevel show the minimum interval between peaks and background noise level. Note that the i-th peak $\boldsymbol{p}(i)$ consists of its amplitude $v(i)$ and time $t(i)$. After that, a sequence of peaks in each channel is precisely searched according to $\boldsymbol{P}$. For example, the i-th peak in channel m, $\boldsymbol{p}_m(i)$, is searched as a peak with maximum value around $t(i)$. Finally, sound events are extracted as

$$\boldsymbol{E} = \{\boldsymbol{e}(i)|1 \leq i \leq N\}, \quad (3)$$

where

$$\boldsymbol{e}(i) = \{\boldsymbol{p}_m(i)|1 \leq m \leq M\}. \quad (4)$$

2.2. Feature Extraction

A 41-dimensional feature vector is extracted, which consists of four types of feature sets; time-domain, spectral, Cepstral and geometrical feature sets shown in Tab. 1. The first three feature sets are extracted using a channel with maximum peak amplitude. The time-domain feature set includes six features. Five of them, *i.e.*, peak amplitude, peak power, time to the next event, attack time, and decay time, are defined in Fig.2, and the last one is a zero cross rate which is defined as the number of zero crossing for the non-rectified signal.

Table 1. Features for Classification

Feature set	Multi-channel	Single-channel
	Used a channel with max pow. 41-dimentional vector	Used fixed channel 37-dimentional vector
Time-domain (6 features)	peak amplitude, peak power, time to the next event, attack time, decay time, zero cross rate	
Spectral (7 features)	amplitudes and frequencies for 1st – 3rd largest peaks, number of peaks	
Cepstral (24 features)	12-dim MFCC, 12-dim Δ MFCC	
Geometrical (4 features)	sound position (x, y, z) sound veclocity (v)	N/A

The spectral feature set includes seven features. Six of them correspond to a local peak in a spectrum of a sound event. Three sets of peak amplitude and its frequency are extracted for the largest to the third largest peaks. The last one is the number of local peaks in the spectrum.

The Cepstral feature set consists of *Mel Frequency Cepstrum Coefficients (MFCC)*. A 24-dimensional feature vector containing 12-dimensional MFCCs and 12-dimensional Δ MFCCs is extracted.

The geometrical feature set includes a sound source position in the Cartesian coordinates and sound source velocity. The three-dimensional sound source position $\boldsymbol{x}_s$ is estimated using 8-channel audio data by assuming that sound power is inversely proportional to the square of the distance from a sound source.

$$\boldsymbol{x}_s(i) = \underset{\boldsymbol{x}}{argmin}\left(\sum_m \left|\frac{a}{(|\boldsymbol{x} - \boldsymbol{x}_m| + r_0)^2} - \boldsymbol{v}_m(i)\right|\right), \quad (5)$$

where $\boldsymbol{x}_m$ is the position of the m-th microphone, a and r_0 are constant values obtained empirically.

This method does not use time or phase difference such as *Time Delay of Arrival (TDoA)*. Rough inter-channel synchronization to find $\boldsymbol{p}_m(i)$ for all channels for the i-th sound event $\boldsymbol{e}(i)$ is sufficient. Therefore, a special multi-channel A/D is not always necessary. The sound source velocity is computed with equations of motion from five temporally-consective positions estimated by Eq. (5).

2.3. Classification with SVM

For classification, we simply used a *Support Vector Machine (SVM)*. After considering four types of sound sources, multi-class SVM with soft margin was selected. The commonly-used *Radial Basis Function (RBF)* was chosen for the kernel. A penalty parameter for soft margin and a kernel parameter for RBF are optimized through two-fold cross validation for training data. As for implementation, libsvm-3.14 was used.

We analyzed the multi-channel feature vector using principal component analysis. Fig. 3 shows that 33 components are necessary to represent 95% of the feature space. Dimension reduction was not performed before classification and the 41-dimensional multi-channel feature vector was directly used as an input for SVM.

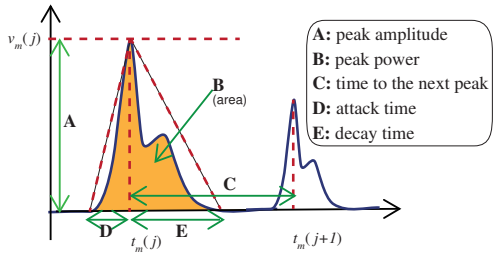


Fig. 2. Time-domain Features

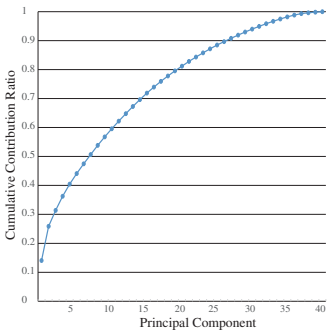


Fig. 3. PCA Result for Multi-channel Features

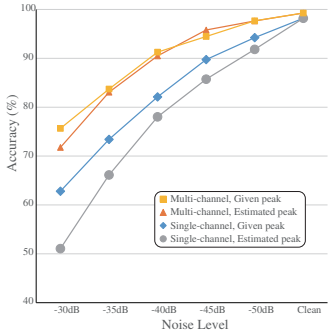


Fig. 4. Classification Result with Single and Multi-channel Features

3. EVALUATION

We evaluated the system to validate the effectiveness of multi-channel features compared to single channel features, and noise-robustness of footstep detection and classification. The differences between four types of feature sets are also discussed.

3.1. Dataset preparation

We recorded four types of sound events, *i.e.*, walking, running, handclaps and utterances, which were made by six persons (male, 20-25 years old). They wore their usual footwear. The floor was covered with a fabric carpet commonly used in an office. For walking and running, we asked each person to follow a gray arrow as shown in Fig. 1, that is, going from $(0.6\,m, 0\,m)$ to $(0.6\,m, 3.6\,m)$ and back to the starting point. For the handclaps, they were asked to make handclaps at $(0.6\,m, 0\,m)$. For the utterances, we asked two persons to simulate a conversation. One was located at $(0.6\,m, 0\,m)$, and the other was at $(0\,m, 1.0\,m)$. Note that no overlapping utterance was made in this case. The number of sound events for each type was around 100 for each person. Therefore, in total, approximately 600 sound events for each type were recorded.

For comparison, we also evaluated system performance based on a single-channel feature vector shown in Tab. 1. It is a 37-dimensional vector consisting of time-domain, spectral and Cepstral feature sets. Since only one microphone was used, geometrical information was unavailable. For feature extraction of the single channel feature vector, it is impossible to take a coarse-to-fine approach with a single microphone, and thus we simply applied ***findpeaks*** for $|s_m(n)|$, not for $s(n)$ in Eqs. (1) and (2).

To investigate noise robustness of the system, we generated noise-contaminated sound events. We added white noise to the recorded sound events because white noise has a wide spectrum and thus it is the hardest noise to be dealt with. The white noise level was changed from -30 dB to -50 dB by 5 dB intervals. The white noise level of 0 dB means that its amplitude is the maximum value in the waveform. In our recorded data, The white noise level of -30 dB almost corresponds to 0 dB in *Signal-to-Noise Ratio (SNR)*.

3.2. Experimental Result and Discussions

We performed two experiments on sound classification as follows:

1. Classification in different noise levels with single-channel and multi-channel feature vectors.
2. Classification for each feature type and each sound event type for more detailed analysis.

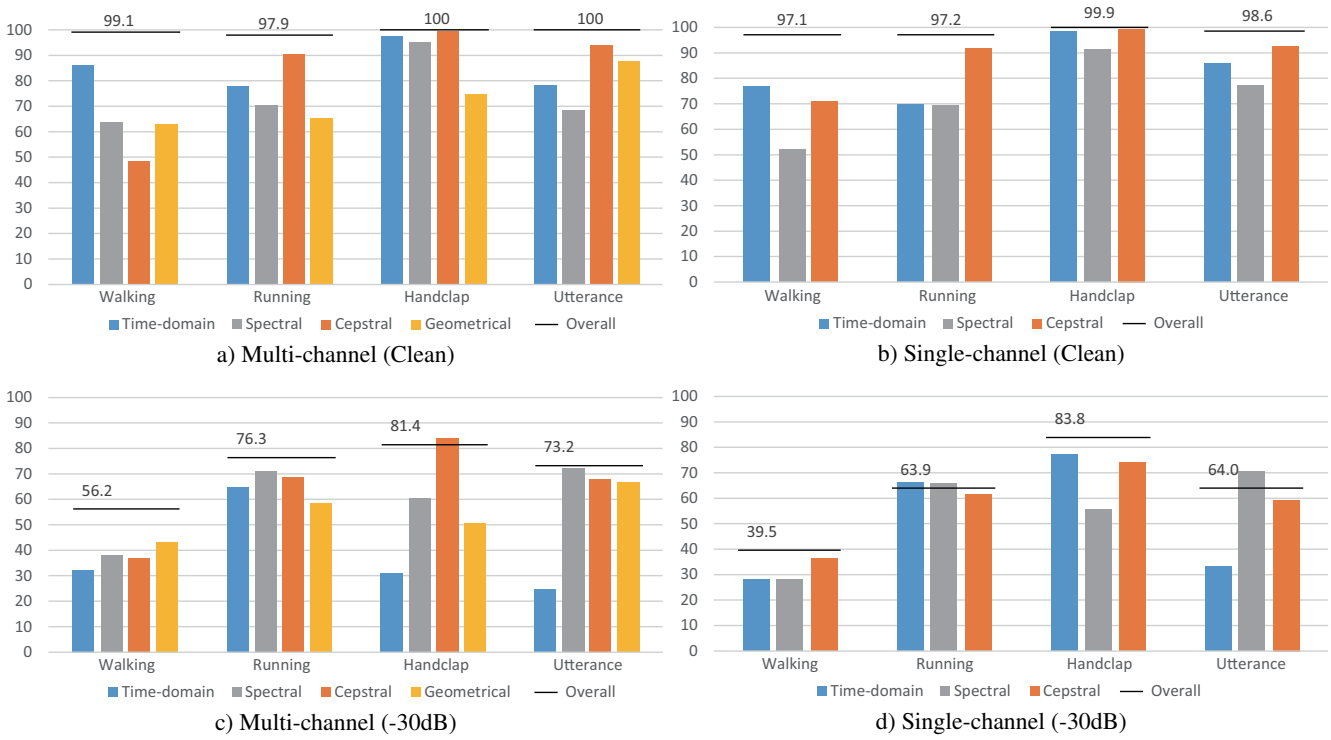
In all cases, SVM was trained with clean data, and test data did not include any training data, *i.e.*. open test.

Fig. 4 depicts the result of the first experiment. The horizontal axis shows white noise levels, and "Clean" signifies that no white noise was added. The vertical axis shows classification accuracy averaged over the four types of sound events. The lines with Multi-channel and Single-channel are the corresponding results for multi-channel and single-channel feature vectors, respectively. "Estimated peak" means that sound event detection was activated, and "Given peak" means that the system used sound event detection result with multi-channel clean data for all noise conditions.

Fig. 5 illustrates the classification accuracy for each feature and event type. Fig. 5a) and b) shows the results with "Clean" data and Fig. 5c) and d) shows those with sound data which -30 dB white noise was added to. Each bar shows the classification accuracy when only a time-domain, Cepstral, spectral, or geometrical feature set was used. A black line with a value shows the classification accuracy when all the feature sets are used together.

From Fig. 4, the use of the multi-channel feature vector is more noise-robust than the single-channel. The difference in classification accuracy is getting larger as the input sound becomes noisier. Without any noise reduction like sound source separation, the system maintained over 70% accuracy with the multi-channel feature vector for -30 dB white noise corresponding to 0 dB in SNR. We hypothesized that this is caused by sound event detection because footstep sounds are difficult to be captured with sufficient power with single channel features when a person is distant away from a microphone, particularly, in noisy conditions. Then, we computed the results with "Given peak." The difference in classification performance between "Given peak" and "Estimated peak" was small with the multi-channel feature vector. On the other hand, with the single channel feature vector, the difference was larger, and over 10 points in -30 dB white noise. This shows that sound event detection is problematic in a noisy condition for the single-channel feature vector. Another 10 point difference is existent between the multi-channel and the single-channel, which means that feature extraction with the multi-channel feature vector is also noise-robust compared to the single-channel.

Fig. 5 gives us more detailed information. In most cases except for handclap in Fig. 5c), and running and utterance in Fig. 5d), classification performance with the integrated feature vector marked with "Overall" outperforms that with a feature type (time-domain, spectral, Cepstral, or geometrical). This suggests that a kind of synergy effect by integration occurs rather than selection of the best feature set. The two exceptional cases (handclap in Fig. 5c) and utterance in Fig. 5d)) were caused by the large difference in performance between feature types. In both cases, actually, the performance with the time-domain feature set was extremely low compared to those



a) Multi-channel (Clean)

b) Single-channel (Clean)

c) Multi-channel (-30dB)

d) Single-channel (-30dB)

Fig. 5. Classification Result by Feature and Sound Event Type

with other feature types, and thus we considered that poor performance with the time-domain feature set affected that with the integrated feature vector. In the other case, *i.e.*, running in Fig. 5d)), the integrated feature maintained almost the same performance as other feature types, but we should analyze the reason that it did not improve the performance.

Fig. 5a) and b) shows that all feature types can be widely applicable to general sound classification besides footsteps. By comparing Fig. 5c) and d), the multi-channel approach improves the noise-robustness for time-domain, spectral and Cepstral feature sets. This is because extraction of these features was conducted by selecting a channel with the maximum sound power in the multi-channel approach, while a fixed channel was always used in the single channel approach. By comparing Fig. 5a) and c), the time-domain feature set is easily affected by noise, while spectral and Cepstral feature sets showed high noise-robustness. The use of the geometrical feature set is also effective, it is noise-robust such as spectral and Cepstral feature sets. In particular, walking footstep detection in Fig. 5c), the geometrical feature set has the best score. This shows that multi-channel features have two advantages in improving noise-robustness, *i.e.*, compensation of missing or ambiguous single channel information by selecting the best channel, and providing geometrical information of a sound source.

4. CONCLUSION

We presented footstep detection and classification using multiple microphones distributed on the floor. We proposed to use time-domain, spectral, Cepstral and geometrical feature sets, and for the geometrical feature set, we developed an amplitude-based localization method instead of using conventional microphone array techniques, which makes the system free from the use of a special multi-channel A/D device. We confirmed that multi-channel features are noise robust, since it can provide the best information among all channels, and also geometrical information on a sound source. We also showed that all of four feature sets are widely applicable to classification of other sound sources besides footsteps. Construction of an online classification system and dealing with overlapping sound sources remain as future work.

5. REFERENCES

[1] D. Rosenthal and H. G. Okuno, Eds., *Computational Auditory Scene Analysis*, Lawrence Erlbaum Associates, Mahwah, New Jersey, 1998.

[2] M. Tanaka and H. Inoue, "A study on walk-recognition by frequency analysis of footsteps," *Trans. IEE of Japan*, vol. 119-C, no. 6, pp. 762–763, 1999.

[3] B. She, "Framework of footstep detection in in-door environment," in *15th Int'l Congress on Acoustics (ICA)*, 2004, pp. Mo.P2.1, I–715–I–718.

[4] Y. Shoji *et al.*, "Personal identification using footstep detection," in *Int'l Symposium on Intelligent Signal Processing and Communication Systems (ISPACS)*, 2004, pp. 43–47.

[5] S.G. Iyengar *et al.*, "On the detection of footsteps based on acoustic and seismic sensing," in *41st Annual Asilomar Conference on Signals, Systems and Computers (ACSSC 2007)*, 2007, pp. 2248–2252.

[6] A. Itai and H. Yasukawa, "Footstep classification using simple speech recognition tecnique," in *IEEE Int'l Symposium on Circuits and Systems (ISCAS)*, 2008, pp. 3234–3237.

[7] T. Saitoh *et al.*, "A real-time footstep tracking for monitoring system," in *International Conference on Signal and Image Processing (SIP)*, IASTED, Ed., 2006, pp. 403–408.

[8] M. Shoji, "Passive acoustic sensing of walking, 5th international conference on intelligent sensors," in *Sensor Networks and Information Processing (ISSNIP)*, IEEE, Ed., 2009, pp. 219–224.

[9] R. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE Transactions on Antennas and Propagation*, vol. 34, no. 3, pp. 276–280, 1986.