ISR TECHNICAL REPORT 2008-17

Person Identification and Gender Recognition from Footstep Sound using Modulation Analysis

DeLoney, Chasity, Advisor: Mesgarani, Nima,
Advisor: Fritz, Jonathan

PERSON IDENTIFICATION AND GENDER RECOGNITION FROM FOOTSTEP SOUND USING MODULATION ANALYSIS

Chasity DeLoney, Nima Mesgarani, Jonathan Fritz

ABSTRACT

We describe a person identification system that is based on classifying the sound of the footstep. The classification is done on the spectrotemporal modulations of sound that are estimated using a model of auditory processing. We describe how different footsteps form a unique footprint in the spectrotemporal modulation domain and how this representation captures the user specific signatures. Using this representation, we achieved higher than 60% accuracy in identifying 9 people with three different shoes and two floors. The study demonstrates the efficacy of the spectrotemporal features in the tasks examined.

INTRODUCTION

Automated person identification is an important component of surveillance. An effective approach to person identification is to reduce it, to the problem of identifying physical characteristics. However, in surveillance applications the usual biometrics, such as the fingerprint or the iris, are no longer applicable. Instead, a useful biometric is the walking style of an individual, because it is non-intrusive and can be detected and measured where other methods, based on visual cues fail, such as in darkness. Furthermore, the walking style of an individual is more difficult to disguise as opposed to the static appearance features such as an individual's face.

The fact that footstep sounds convey information about the identity of a person has been shown by [1]. They studied how well people can identify recorded walking sounds of their co-workers on the base of their everyday experience. The results show that a group of 13 people were able to achieve an identification rate of 66%. Researchers have also shown by subjective tests that the footsteps carry gender information [2] and people can recognize the gender with more than 70% accuracy. Automated user identification systems have been developed in the past. For example, an automated gender recognition method for counting store customers has been proposed in [3] which works based on human silhouette shape, footstep and pressure characteristics. Using only footsteps, the authors reported the accuracy of 80%.

Smart Floor [4] and Active Floor [5] are two other systems developed for automated user identification based on walking patterns. They identify users using a set of load cells arranged on the floor tile to measure Ground Reaction Force (GRF). Smart Floor has been able to achieve a recognition rate of 93% for 15 subjects; Active Floor has achieved 91% accuracy. Another user identification system, ubiFloor [6], tracks the user's location with 144 low-cost ON/OFF switch sensors to identify users based on their walking patterns. Experimental results showed that this system can identify the 10 registered users at the rate of 92%. None of the systems described here use the sound of the footstep for person identification. In this study, we describe a method which relies on the unique pattern that the footstep sounds form in spectrotemporal domain: *spectrotemporal footprint*. The features are extracted by a model of the sound processing in the auditory cortex [7] [8]. The sound is recorded using a regular microphone that eliminates the need for expensive pressure sensors. The recorded sounds are then mapped to a high-dimensional modulation representation by an auditory model and then classified by a set of parallel

Support Vector Machines (SVM). We shall briefly review the auditory model in Section 2 and then present the experimental results and performance evaluation of our proposed system in Section 3.

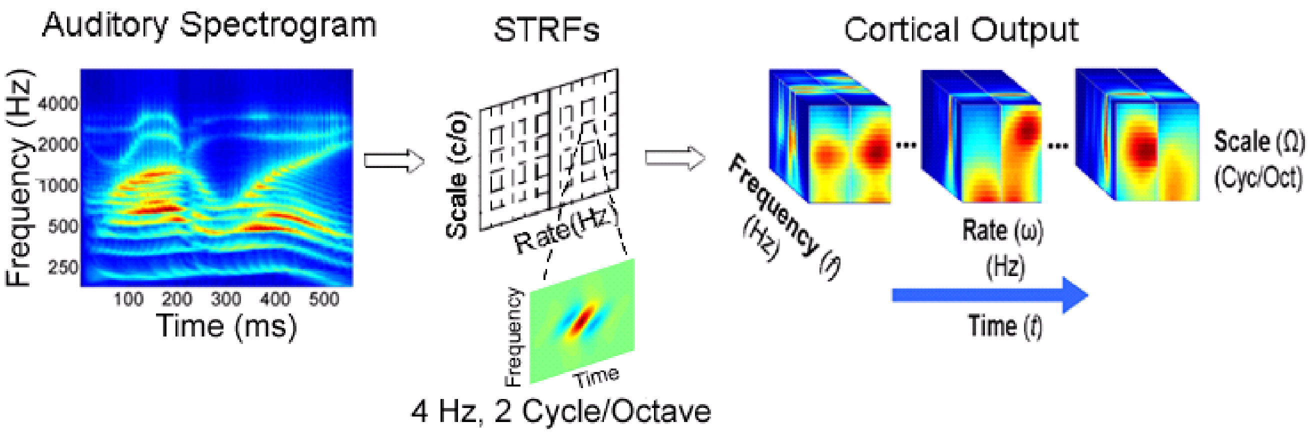

Figure 1. The auditory spectrogram is analyzed by a bank of spectro-temporal modulation selective filters decomposing it into spectrotemporal modulations. The center panel shows one such filter tuned to 4Hz (Rate) and 1 Cycle/Octave (Scale). Total output as a function of time from the model is therefore indexed by three parameters: scale, rate, and frequency.

2. MODULATIONS FEATURES

The spectrotemporal modulation features are extracted from an auditory model inspired by psycho-acoustic and neurophysiologic findings in the early and central stages of the auditory pathway [8].The early stage converts the sound waveform into an *auditory spectrogram* - roughly akin to a time-frequency distribution along a tonotopic (logarithmic frequency) axis [8]. The next stage, the cortical stage, performs a two dimensional wavelet transform of the auditory spectrogram, thus providing an estimate of its spectral and temporal modulation content. It is computationally implemented by a bank of two-dimensional, (spectro-temporal) filters that are selective to different spectro-temporal modulation parameters, which range from slow to fast *rates* temporally (in Hertz), and from narrow to broad *scales* spectrally (in Cycle/Octave). The basic mathematical formulation of the model can be summarized as followed (figure 1):

$$ycochlea(t, f) = s(t) \ast hcochlea(t, f) \quad (1)$$

$$yan(t, f) = gcochlea(\partial tycochlea(t, f)) \ast \mu haircell \ (2)$$

$$y(t, f) = max(\partial f\, yan(t, f), 0) \ast \mu midbrain \ (3)$$

$$r\pm(t, f; \omega, \Omega) = (4)$$

$$y(t, f) \ast tf\, [STRF\pm(t, f; \omega, \Omega)]$$

where, $s(t)$ is the sound, $ycochlea(t, f)$ is the cochlear filter output, $yan(t, f)$ is auditory nerve patterns, $y(t, f)$ is the auditory spectrogram and $r\pm(t, f; \omega, \Omega)$ is the cortical representation. The sign of r specifies the direction of spectrotemporal modulation where − is for downward and + denotes upward patterns. The modulation representation of sound, equation (4) is a 4-dimensional function of time (t), frequency (f), rate (ω) and scale (Ω). If we average the time dimension on a given duration of sound, we obtain the average rate-scale-frequency representation in that given time window, this is then used for the purpose of classification. The output of this stage is normalized and fed into the classification stage which consists of 16 individual classifiers. Each classifier is trained to identify one person from the rest, and the final output is obtained by taking the classifier that has the maximum certainty. Classifiers used were linear Support Vector Machine (SVM) that minimizes the error by finding the optimal hyperplane, maximizing the margin between two classes. The certainty of the output is defined as the distance of each sample from the classifier hyperplane. Since the classifiers are linear, the normal vector of the hyperplane can be used to measure the contribution of each rate, scale and frequency to the decision of the classifier and as a result, to identify the signature of each person in this domain (section 3.5).

3. EXPERIMENTAL RESULTS

3.1. Data Collection

The footstep data was collected using a laptop and a general purpose microphone taped to the floor. The sounds were recorded with actual noise in background while subjects were walking normally.

The sound data were then filtered with a high-pass filter with cut-of frequency of 125Hz to reduce the recording noise generated by the computer. Footstep sounds were recorded from 9 subjects, wearing three different kinds of shoes and on two floors (tile and wood). The amount of data gathered for each floor/shoe/subject condition was 120 seconds. Each subject brought two different types of shoes to the recording session; the third shoe was the same for all subjects. We used 90% of the data for training and 10% for testing; 50 random training-test subsets were produced and we measured the average classification across subsets for accuracy analysis.

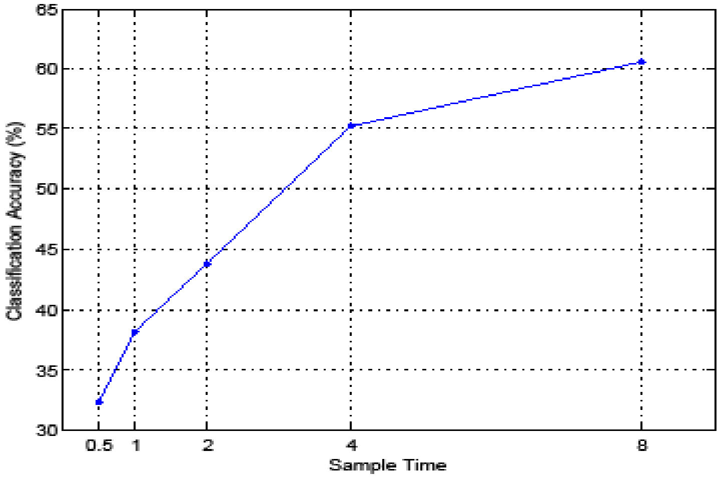

Figure 2. Percentage of correctly identified people as a function of the duration of footstep sound. As shown in this figure, increasing the duration of the sound also increases the accuracy of identification substantially.

3.2. Effect of time length

One of the parameters that directly affect the classification accuracy is the duration of the footstep signal that is used for classification. The longer the sound, the more accurate the estimation of the average spectrotemporal modulation is which, in turn, results in improved classification accuracy. Figure 2 shows the overall percentage of correctly classified samples versus the duration of footstep sound from 0.5 to 8 seconds. The figure shows the accuracy of identification improves substantially as the duration of the footstep sound increases, starting from 32% at 0.5 second to 60.5% at 8 seconds. For the rest of the analysis in this study, we fixed the duration of sound to four seconds where the identification rate is 55%.

3.3. Person identification

Figure 3 demonstrates the confusion matrix for 9 subjects when the duration of footstep sound is 4 seconds and the training and testing is done for all the combination of all shoes. The rows indicate the actual identity of the subjects and the columns correspond to the identified one by the classifiers. As Figure 3 shows, some subjects were more easily confused than the others as can be seen from the groups formed in the figure.

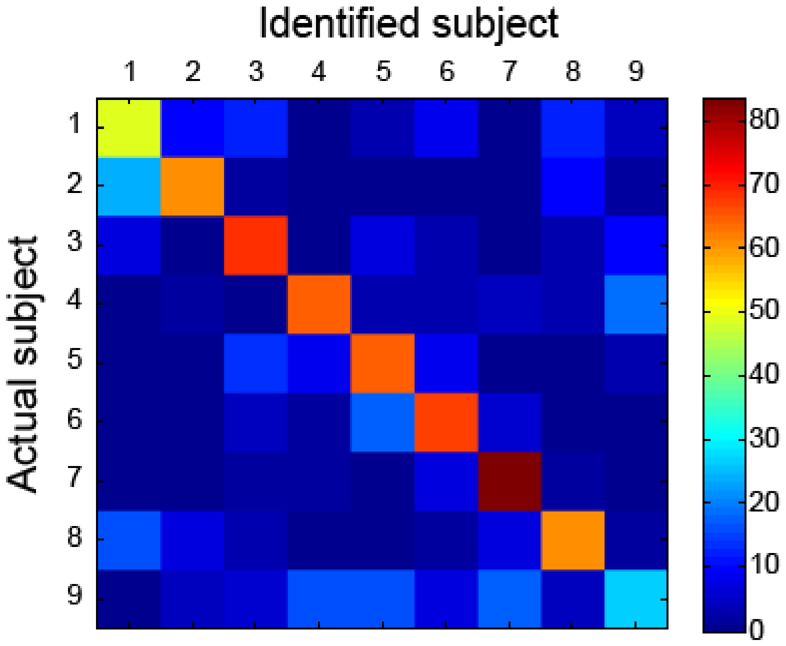

Figure 3. Confusion matrix for 9 subjects. The classifiers were trained and tested on all three shoes.

3.5. Generalization to novel shoes

We ran a series of tests to test the generalization of the system to unseen shoes and floors. In the first series of tests, we trained the system on one shoe, and test it on other ones. These tests reveal the robustness of the system to shoe change, which is an important characteristic of an identification system that is based on footstep. These tests are summarized in

Table 1, Table 2, and Table 3.

Table 1. Confusion matrix for different shoes, floor 1

	Test Shoe 1	Test Shoe 2	Test Shoe 3
Train Shoe 1	70%	14%	23%
Train Shoe 2	11%	88%	15%
Train Shoe 3	18%	24%	46%

Table 2. Confusion matrix for different shoes, floor 2

	Test Shoe 1	Test Shoe 2	Test Shoe 3
Train Shoe 1	78%	28%	32%
Train Shoe 2	31%	77%	20%
Train Shoe 3	36%	16%	61%

Table 3. Confusion matrix for different shoes, floor 1 and 2 combined

	Test Shoe 1	Test Shoe 2	Test Shoe 3
Train Shoe 1	66%	19%	21%
Train Shoe 2	13%	74%	16%
Train Shoe 3	25%	20%	39%

As these tables show, the overall generalization of the system to unseen shoes is not very good, and for this system to work, it needs to be trained on the shoes. We observed the same trend for classifiers that were trained on two of the shoes, and were tested on the third one. Again, the system performed significantly better on the shoes that it had been trained on (

Table 4, Table 5, Table 6).

Table 4. Confusion matrix for classifiers trained on two shoes and tested on the third one, floor 1

	Test Shoe 1	Test Shoe 2	Test Shoe 3
Train Shoe 1 & 2	51%	79%	20%
Train Shoe 1 & 3	65%	15%	40%
Train Shoe 2 & 3	23%	81%	34%

Table 5. Confusion matrix for classifiers trained on two shoes and tested on the third one, floor 2

	Test Shoe 1	Test Shoe 2	Test Shoe 3
Train Shoe 1 & 2	70%	65%	34%
Train Shoe 1 & 3	69%	23%	55%
Train Shoe 2 & 3	36%	67%	48%

Table 6. Confusion matrix for classifiers trained on two shoes and tested on the third one, floors 1 & 2

	Test Shoe 1	Test Shoe 2	Test Shoe 3
Train Shoe 1 & 2	49%	61%	21%
Train Shoe 1 & 3	57%	18%	34%
Train Shoe 2 & 3	19%	66%	28%

3.6. Generalization to a novel floor

We also tested the ability of the system trained on a specific floor to a new floor. The generalization characteristics of the system is important and determines whether the system has to be trained on the same floor that it is going to be trained on. To test this, we trained the system on one of the floors and tested them on the other one. We ran these tests for each shoe and all the shoes combined as shown in Table 7, Table 8, Table 9, and Table 10.

Table 7. Confusion matrix across floors, shoe one

	Test Floor 1	Test Floor 2
Train Floor 1	70%	31%

Train Floor 2	31%	75%

Table 8. Confusion matrix across floors, shoe two

	Test Floor 1	Test Floor 2
Train Floor 1	88%	28%
Train Floor 2	35%	80%

Table 9. Confusion matrix across floors, shoe three

	Test Floor 1	Test Floor 2
Train Floor 1	86%	22%
Train Floor 2	24%	49%

Table 10. Confusion matrix across floors, all shoes

	Test Floor 1	Test Floor 2
Train Floor 1	50%	19%
Train Floor 2	21%	54%

3.7. Spectrotemporal footprint

As shown in previous section, it is possible to identify people based on the sound of their footstep. This suggests different people, based on how they walk, what they wear and the type of the floor , create a different *footprint* in this representation. In this section we examine the modulation features more thoroughly to provide better insight on how this system works. Figure 4 shows the average rate-scale display for 9 subjects for rate from -32 to 32Hz and scales from 0.5 to 8Cyc/Oct. Each row corresponds to a specific floor and shoe combination. Red and blue indicate presence and absence of modulations in corresponding rate and scale. The footprints for the first floor are shown in the top three rows of Figure 4.

Rows one, two, and three correspond to three different shoes. The next three rows are the footprints on the second floor. We observe that the first floor generates more high-scale energy than the second floor (concentrated at the top of panels for floor one, and on the bottom of panels for floor two). We think this has to do with the shape of the frequency profile that is generated by each floor. Another assumption based on the results we received is that some subjects have less variability across different shoes, for example subjects 6 and 7 produce footprints that are not inconsistent across shoes, however some subjects, (Subject 1 for example) produce different patterns for different shoes. In short, figure 4 shows how variations in the subject, the floor and the shoe, change the representation of the footstep sound in the modulation domain. Depending on the task, the classifiers is trained to learn the invariant dimension and ignore the irrelevant variability.

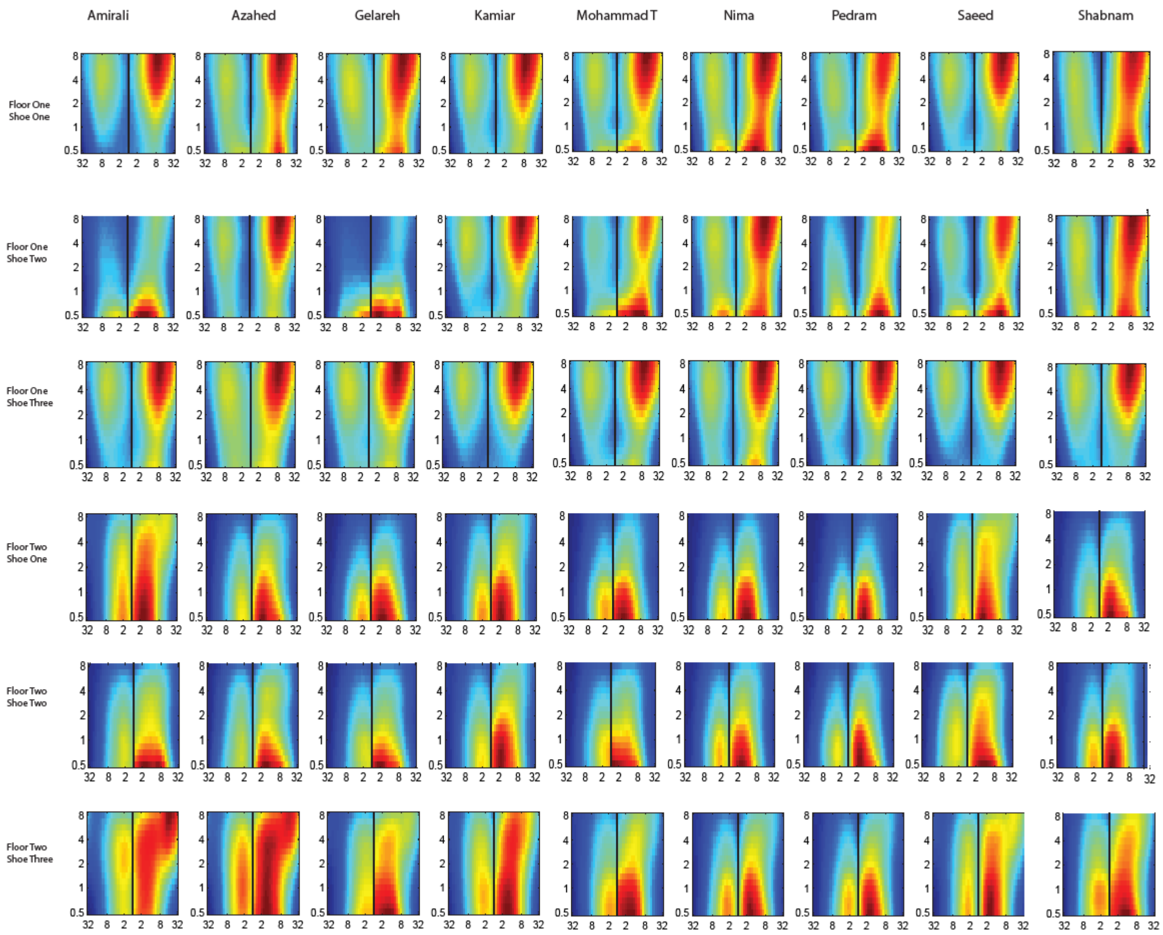

Figure 4. Spectrotemporal footprints. Average spectrotemporal modulations for different subjects, with three different shoes and on the two floors. Top three rows correspond to the first floor and the bottom three to the second one.

4. SUMMARY AND CONCLUSION

A footstep identification algorithm is described based on the spectrotemporal modulations of sound extracted from a model of the auditory system. The performance of the system was tested on a 9 people identification task with three different shoe and two floors and the dependence of the accuracy on the duration of sound was examined. We showed the accuracy as high as 88% for one shoe, one floor and 54% for a system trained on three shoes. This shows that this modulation representation is capable of capturing the *footprint* of the way different people walk. We would like to study ways to improve the performance of the system, by investigating other features that can be estimated from sound to find out their importance for classification. In addition, we would like to carry on psychoacoustical studies to measure how well humans can perform the same task, and measure their generalization to novel shoes and floors.

6. REFERENCES

[1] K. Makela, J. Hakulinen, M. Turunen, ”The use of walking sounds in supporting awareness, ”Proceedings of the 2003 International Conference on Auditory Display, Boston, MA, USA, 2003.

[2] X. Li, R.J. Logan, R.E. Pastore, ”Perception of acoustic source characteristics: Walking sounds,” Journal of the Acoustical Society of America, Vol. 90, pp. 3036-3049, 1991.

[3] K. Sudo, J. Yamato, A. Tomono, K.I. Ishii, ” Gender recognition method based on silhouette, footstep, and foot pressure measurements for counting customers,” Electronics and Communications in Japan, Part 2, Vol. 85, No. 8, pp. 54-64, 2002.

[4] R. Orr, G. Abowd, ”The smart floor: A mechanism for natural user identification and tracking,” Proceedings of the Conference on Human Factors in Computing Systems , Hague, Netherlands, pp. 1-6, 2000.

[5] M.D. Addlesee, A.H. Jones, F. Livesey, and F.S. Samaria, ”ORL active floor,” IEEE Personal Communications, Vol. 4, No. 5, pp. 35-41, 1997.

[6] J.S. Yun, S.H Lee, W.T.Woo, and J.H. Ryu, ”The user identification system using walking pattern over the ubifloor,” in Proceedings of the International Conference on Control, Automation, and Systems, Gyeongju, Korea, 2003.

[7] N. Mesgarani, M. Slaney, S. A. Shamma, ”Speech discrimination based on multiscale spectro-temporal features”, IEEE International Conference on Acoustic, Speech and Signal Processing, Montreal, Canada, 2004.

[8] T. Chi, P. Ru, and S.A Shamma, ”Multiresolution spectrotemporal analysis of complex sounds,” Journal of the Acoustical Society of America, 2005.