Speaker Identification using Three Signal Voice Domains during Human-Robot Interaction

[Late Breaking Reports]

Fernando Alonso-Martín
Universidad Carlos III Madrid
Avenida de la Universidad, 30
Leganés, Madrid (Spain)
famartin@ing.uc3m.es

Arnaud Ramey
Universidad Carlos III Madrid
Avenida de la Universidad, 30
Leganés, Madrid (Spain)
arnaud.a.ramey@gmail.com

Miguel Angel Salichs
Universidad Carlos III Madrid
Avenida de la Universidad, 30
Leganés, Madrid (Spain)
salichs@ing.uc3m.es

ABSTRACT

This LBR describes a novel method for user recognition in HRI, based on analyzing the peculiarities of users voices, and specially focused at being used in a robotic system. The method is inspired by acoustic fingerprinting techniques, and is made of two phases: a)*enrollment in the system*: the features of the user's voice are stored in files called *voiceprints*, b)*searching phase*: the features extracted in real time are compared with the voiceprints using a pattern matching method to obtain the most likely user (match).
The audio samples are described thanks to features in three different signal domains: time, frequency, and time-frequency. Using the combination of these three domains has enabled significant increases in the accuracy of user identification compared to existing techniques. Several tests using an independent user voice database show that only half a second of user voice is enough to identify the speaker. The recognition is text-independent: users do not need to say a specific sentence (key-pass) to get identified for the robot.

Keywords

Acoustic fingerprint, speaker identification, feature extraction, patter matching

1. INTRODUCTION

In order to achieve a natural and personalized interaction between a robot and its human users (HRI), it is important that the former identifies correctly its human peers. Audio features are often used to reach that point.
Previous works by other authors ([5]) include several descriptions of systems enabling a caller to obtain access via telephone network to services by entering a spoken key-pass (text-dependent method). In order to achieve this goal, in the enrollment phase, first automatic speech recognition (ASR) identifies the key-pass; then a voice verification algorithm is used. Some features of the voice are extracted and stored next to the key-pass sentence. In the verification phase, real time features are compared with the previously stored features. The identification is considered as successful if the compared distance between both sets of features is below a given threshold (pattern matching approximation using text-dependent verification).
A review of acoustic fingerprint methods [3] report their use for a wide range of applications, such as identifying songs, contents, etc. However, their use for the identification of the user is more challenging, because two identical sentences uttered by the same person result in two different signals. Our proposed architecture tends to overcome these difficulties.

HRI'14, March 3–6, 2014, Bielefeld, Germany.
ACM 978-1-4503-2658-2/14/03.
http://dx.doi.org/10.1145/2559636.2563706.

2. DESCRIPTION OF THE SYSTEM

The previous system used by the robot Maggie ([2]) required the user to enroll on the system. This phase consisted of questions asked by the robot to the user. Along with her name, age, language, a key-pass sentence is chosen by the user to learn her voice tone. It can be a digit password or anything else. For this task, we used the third-party Loquendo ASR-Speaker Verification package [4]. The main drawback was that, to be correctly identified, the user needed to utter the same sentence that he used in the enrollment phase.

System overview. In order to avoid this drawback, we implemented a new module of *text-independent* user identification based on *pattern matching* techniques. During a similar question-based (but key-pass free) enrollment phase, the robot learned features of the voice of the new user. These features are stored into a so-called voiceprint file. Details about feature extraction and matching come below.
The work presented in this paper, includes user identification using voice features. It is implemented in Chuck and C++. It is part of a complete multimodal interaction system applied to HRI called *Robotics Dialog System* ([1]). This RDS interaction system is based on the ROS architecture [1] and supplies the interaction capabilities for the RoboticsLab robots [2].

Used features. Two aspects of the new system can be underlined: the features are extracted without needing a specific key-pass phase, and these extracted features belong to

[1]http://wiki.ros.org
[2]http://roboticslab.uc3m.es/

three different domains: time, frequency, and time-frequency (more details about them in [1]). The used features are *Root Mean Square (RMS)* (computed on time domain); *Pitch computed using Fast Fourier Transform* (frequency domain); *Pitch computed using Haar Discrete Wavelet Transform* (time-frequency domain); *Flux* (frequency domain); *RollOff* (frequency domain); *Centroid* (frequency domain); *Zero-crossing rate (ZCR)* (time domain).

Real time feature extraction. Once the voiceprints of the enrolled users have been generated and stored, it is possible to identify which user is speaking at any time. *Voice Activity Detection*, which consists in detecting whether users are talking or quiet, is based on our previous work [1]. When voice activity is detected, the previously listed audio features are extracted. Each second of voice generates several rows of features. Each row is called a *bit*, and a set of *bits* is known as *sub-fingerprint*. The number of bits that compose a sub-fingerprint will be called *sub-fingeprint size*, and is the minimum information to identify an user.

Our system works with a sample rate of 44100 Hz and a window size of 4096 samples, therefore each bit of fingerprint corresponds to 0.1 second of voice (i.e. 1 second of voice generates 10 bits).

Identification (matching) phase. The identification is made by computing the distance between the current audio features and the features of the enrolled users, stored in each of the voiceprint files. If the best-match distance is below a threshold, the speaking user is considered to be the user that generated that voiceprint file.

In mathematical terms, let $n = 7$ the number of different audio features. The distance d between a current *bit* C of n extracted features and a voice print file V, a set of n_v bits of n features, is computed with the following formula:

$$d(C, V) = \min_{j \in [0, n_v]} \|[1 \dots 1] - \exp(-\alpha \circ |C - V_j|)\|_{L_1}$$

where $\|.\|_{L_1}$ is the Manhattan distance, $\circ$ is the element-wise product, and $\alpha = [\alpha_0 \dots \alpha_n]$ a scaling vector.
The best match for a given bit C is then defined as the voice print which bits obtains the minimum distance with C. It is considered a valid match if that distance is below an empirical threshold.
The distance computation presented above is called *bit-to-bit* ($n \times 1$ vectors comparison). It can also be made on successive sets of *bits* ($n \times w, w \in \mathbf{N}$ matrices comparison, w is called the sub-fingerprint size).

3. EXPERIMENTAL RESULTS

In order to verify the system accuracy to identify human peers, we have used a multi-language voice database [3] Some samples were used for generating the voiceprints, and other for evaluating if their best match corresponded to the correct user.
Figure 1 that relates the number of enrolled users, and the sub-fingerprint size, to the accuracy rate of the user identification. The accuracy rate is close to 100% for few users, and while it decrease while the number of users increase, it remains over 70% for as many as eighteen users for the best choice of sub-fingerprint size. The best results are obtained

[3]The database used was VoxDB (http://voxdb.org).

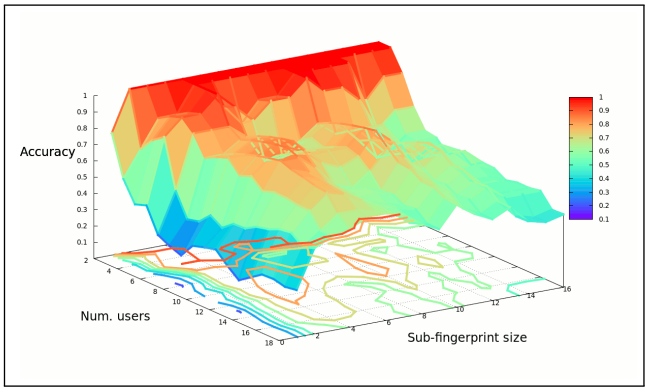

Figure 1: Number of Users vs Accuracy vs Sub-fingerprint size

using a sub-fingerprint size of 5, i.e. a voice sample of half a second is long enough to identify the speaker. Similar work, presented by Lite[6], claims an accuracy rate of 79% for 11 users and 3 seconds for each identification. Our system, with 11 users gets an accuracy rate of 86% using 0.5 seconds of voice.

4. CONCLUSION AND FUTURE WORK

In this paper, we have presented a real-time user recognition algorithm based on acoustic fingerprint methods and tailored for HRI. Audio features computed in three acoustic domains are matched with lightweight pattern matching techniques, which confer the robot an intuitive and accurate user recognition system. Future work will extend the benchmarking of our system against others.

5. ACKNOWLEDGMENTS

The authors gratefully acknowledge the funds provided by the Spanish MICINN (Ministry of Science and Innovation) through the project "Aplicaciones de los robots sociales", DPI2011-26980 from the Spanish Ministry of Economy and Competitiveness. Moreover, the research leading to these results has received funding from the RoboCity2030-II-CM project (S2009/DPI-1559), funded by Programas de Actividades I+D en la Comunidad de Madrid and cofunded by Structural Funds of the EU.

6. REFERENCES

[1] F. Alonso-Martin, A. Castro-González, J. Gorostiza, and M. A. Salichs. Multidomain Voice Activity Detection during Human-Robot Interaction. In *International Conference on Social Robotics (ICSR 2013)*, pages 64–73, Bristol, 2013. Springer International Publishing.

[2] F. Alonso-Martín, J. F. Gorostiza, M. Malfaz, and M. Salichs. Multimodal Fusion as Communicative Acts during Human-Robot Interaction. *Cybernetics and Systems*, 44(8):681–703, 2013.

[3] J. Cano, P., Batle, E., Kalker, T., & Haitsma. A review of algorithms for audio fingerprinting. In *In Multimedia Signal Processing, IEEE Workshop on*, pages 169–173, 2002.

[4] E. Dalmasso, F. Castaldo, P. Laface, D. Colibro, and C. Vair. Loquendo - Speaker recognition evaluation system. In *Acoustics, Speech and Signal Processing, ICASSP 2009. IEEE International Conference on*, pages 4213–4216, Taipei.

[5] T. B. Hunt, Alan K.; Schalk. Simultaneous voice recognition and verification to allow access to telephone network services, 1996.

[6] K. P. Li. Text-independent speaker recognition with short utterances. *The Journal of the Acoustical Society of America*, 72(S1):S29, Aug. 1982.