The International Journal of Speech, Language and the Law

IJSLL (PRINT) ISSN 1748-8885
IJSLL (ONLINE) ISSN 1748-8893

PhD abstract

# Fusing prosodic and acoustic information for speaker recognition

*Mireia Farrús*

*Researcher*
*TALP Research Centre*
*Department of Signal Theory and Communications*
*Universitat Politècnica de Catalunya*
*08034 Barcelona*
*Spain*

Awarding institution: Universitat Politècnica de Catalunya, Spain
Date of award: October 2008

KEYWORDS: SPEAKER RECOGNITION; PROSODY; MULTIMODALITY; IMITATION; CONVERSION

Automatic speaker recognition is the use of a machine to identify an individual from a spoken sentence. Recently, this technology has been increasingly used in applications such as access control, transaction authentication, law enforcement, and system customisation, among others (Campbell 1997).

One of the central questions addressed by this field is what it is in the speech signal that conveys speaker identity. Traditionally, automatic speaker recognition systems have relied mostly on short-term features related to the spectrum of the voice (Rabiner and Juang 1993). However, human speaker recognition relies on other sources of information; therefore, there

**Contact**
email: mfarrus@gps.tsc.upc.edu

IJSLL VOL 16.1 2009 169–171
doi : 10.1558/ijsll.v16i1.169

is reason to believe that these sources can also play an important role in automatic speaker recognition tasks, adding complementary knowledge to the traditional spectrum-based recognition systems and thus improving their accuracy (Peskin et al. 2003).

The main objective of this thesis was to add prosodic information to a traditional spectral system in order to improve its performance. To this end, several characteristics related to human speech prosody – which is conveyed through intonation, rhythm and stress – were selected and combined with the existing spectral features. A preliminary speaker verification system based on prosodic features was then built in order to improve a voice spectrum based verification system over the conversational Switchboard-I database. Whereas the performance of a spectral system in these experiments was considerably improved by using those prosodic features related to segment duration and fundamental frequency, the information contained in the pauses was not useful to improve either of the spectral and the rest of prosodic features.

Furthermore, this thesis focused on the use of additional acoustic features – namely jitter and shimmer – in order to improve the performance of the proposed spectral-prosodic verification system. Both jitter and shimmer features are related to the shape and dimension of the vocal tract, and they have been largely used to detect voice pathologies. The results showed that jitter and shimmer can be used to provide complementary information to both spectral and prosodic systems, and that the absolute measurements of both jitter and shimmer parameters seem to be more speaker discriminant than their corresponding relative measurements.

Since almost all the above-mentioned applications can be used in a multimodal environment, this thesis also aimed to combine the voice features used in the speaker recognition system together with other biometric identifiers in order to improve the global performance (Bolle, Connell, Pankanti, Ratha and Senior 2004). To this end, the speech features were combined with facial features using several normalisation and fusion techniques, and the final fusion results were improved by applying different fusion strategies based on sequences of several steps. The results varied largely depending on the fusion technique utilised. The use of support vector machines, for instance, outperformed the overall fusion results obtained with matcher weighting technique. Furthermore, multimodal fusion was also improved by applying a histogram equalisation to the unimodal score distributions as a normalisation technique

On the other hand, it is well known that humans are able to identify others from voice even when their voices are disguised. The question arises as to how vulnerable speaker recognition systems are against different voice disguises, such as human imitation (Zetterholm 2003) or artificial voice conversion,

which are potential threats to security systems that rely on automatic speaker recognition. The last chapter of the thesis deals with the robustness of speaker recognition to imitated and converted voices, where several experiments were designed in order to test the vulnerability of the speaker recognition task in different imitation environments. The first experiment was performed in order to test the influence of foreign accents and dialects – as a sort of imitation – in auditory speaker recognition, and it showed that dialect imitation could confuse both the human listener and the speaker recognition system. In a second experiment, the voices of two well-known professional imitators to impersonate several well-known politicians were used to analyse the behaviour of some selected and acoustic features in the imitated voices. As a result, the identification error rate for most of the selected prosodic and source-related features increased when testing the mimicking voices with respect to the natural voices.

Finally, some experiments tried to analyse the behaviour of an automatic speaker recognition system in front of automatic converted voices. They showed that most of the converted voices were identified as their corresponding target speaker. However, they failed sometimes to deceive the system and the source voice was recognised, especially in the intragender conversions, which leads to think that the recognition system may be more robust to these kind of conversions that the crossgender ones. The results also revealed that some voices are more difficult to convert than others, and that the correct identification decreases as the amount of conversion training data increases.

## References

Bolle, R. M., Connell, J. H., Pankanti, S., Ratha, N. K. and Senior, A. W. (2004) *Guide to Biometrics.* New York: Springer.

Campbell, J. P. (1997) Speaker recognition: a tutorial. *IEEE* 85: 1437–1462.

Peskin, B., Navrátil, J., Abramson, J., Jones, D., Klusáèek, D., Reynolds, D.A. and Bing Xiang (2003) *Using Prosodic and Conversational Features for High-performance Speaker Recognition: report from JHU WS'02.* Paper presented at the ICASSP, Hong Kong, April.

Rabiner, L. R. and Juang, B. H. (1993) *Fundamentals of Speech Recognition.* Englewood Cliffs, New Jersey: Prentice Hall, Inc.

Zetterholm, E. (2003) *Voice Imitation. A Phonetic Study of Perceptual Illusions and Acoustic Success.* Lund: Lund University.