

Acoustic Features for Profiling Mobile Users of Conversational Interfaces

Dave Toney¹, David Feinberg² and Korin Richmond¹

¹ School of Informatics, University of Edinburgh EH8 9LW.

² School of Psychology, University of St. Andrews KY16 9JP.
{dave, korin}@cstr.ed.ac.uk, drf3@st-and.ac.uk

Abstract. Conversational interfaces allow human users to use spoken language to interact with computer-based information services. In this paper we examine the potential for personalizing speech-based human-computer interaction according to the user's gender and age. We describe a system that uses acoustic features of the user's speech to automatically estimate these physical characteristics. We discuss the difficulties of implementing this process in relation to the high level of environmental noise that is typical of mobile human-computer interaction.

1 Introduction

Conversational interfaces allow human users to use spoken language to interact with computer-based information services. Typically, these interfaces are implemented by integrating speech-processing, natural language, and telecommunications systems [1]. With the addition of location-awareness technology, conversational interfaces can provide mobile users with personalized services [2]. For example, traffic, travel, weather and tourist information can all be enhanced by knowledge of a user's location.

In this paper we examine the potential for personalizing speech-based human-computer interaction according to the user's physical characteristics. Specifically, we focus on two characteristics of the user: gender and age. In a conversational interface, an estimate of these characteristics can be useful for influencing the style and content of computer-generated utterances. Commercially-orientated services, for instance, can make use of gender differences in consumer behaviour and select their content accordingly [3]. Similarly, hearing ability is known to decrease with age. An adaptive interface can use an estimate of a user's age to adjust the volume level of its utterances [4]

In order to profile a user in this way, a number of acoustic features must be extracted from his/her voice. However, extracting these acoustic features in a mobile context is problematic. The high level of background noise associated with the use of mobile (cellular) phones or in-vehicle devices often restricts the performance of systems based on acoustic feature extraction [5]. In this study we investigate whether an estimation of gender and age is possible within a mobile setting, in spite of the associated background noise.

In the following sections we propose a set of acoustic features for estimating speaker gender and age. We then outline an implementation of this estimation process and evaluate its performance. We conclude with a summary of our findings and suggest future directions of research.

2 Acoustic Features for User Profiling

What features of the human voice can be used to differentiate one subset of the population from another? More specifically, what acoustic features can help to distinguish between male and female voices, and between younger and older voices? Previous studies [6, 7] have identified three acoustically-based features for identifying a person’s gender and age: (i) fundamental frequency (F_0), (ii) jitter and (iii) shimmer. A segment of speech contains a range of sound wave frequencies. The fundamental frequency is the lowest frequency component and is perceived as voice pitch. Jitter and shimmer, on the other hand, are associated with much more subtle voice qualities. Jitter relates to the variability of the fundamental frequency while shimmer refers to the variation in amplitude of successive pitch periods. Large amounts of jitter and shimmer are often manifested in voices that sound “shaky” or “trembling”. In addition to these three features, we make use of a fourth, known as harmonics-to-noise ratio (HNR). HNR is a measure of the amount of noise in a speech signal. Although a high level of noise is expected with mobile communications, a *relative* increase in HNR may indicate older or pathological voices [8].

The measurement of these four voice qualities may be compromised by the noisy environments that are often experienced in mobile communication. To illustrate the effect of noise, Figure 1 contains two spectrograms. A spectrogram shows the variation in energy over time of different vocal frequencies. In this case, the phrase “critical component” was recorded once using a microphone and then again using a mobile phone. The spectrograms of these recordings show the energy variations for frequencies in the range 0-8kHz during a time period of just under one second. Even without a detailed understanding of the information contained in the spectrograms, it should be clear that the second recording contains a great deal of noise.

3 Implementation

We now describe how the acoustic features discussed in the previous section were used to implement an automatic estimator of speaker gender and age. The purpose of the estimator program was to input an example of a user’s speech as a .wav sound file and output a gender classification and age estimate.

Previous work on age estimation has concentrated on deciding whether a speaker was elderly or not [7]. Acoustic measures such as jitter and shimmer are known to increase appreciably in the elderly. However, in this study we looked at speakers in the age range 21–55. Our aim was to investigate whether an estimation of age could be achieved in non-elderly speakers.

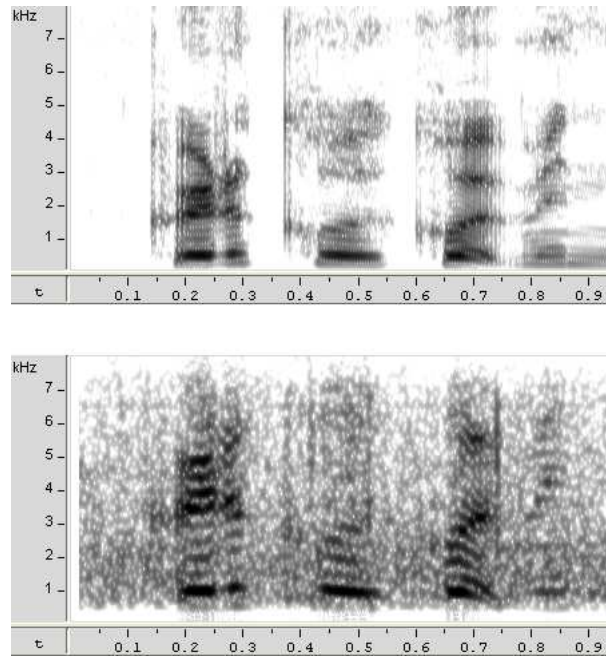


Fig. 1. Spectrograms of the phrase “critical component” as spoken through a microphone (top) and a mobile phone (bottom)

3.1 Tools and Materials

We used two open source tools, Praat [9] and Netlab [10] and one commercial product, Matlab [11]. Praat is an extensive speech analysis and synthesis tool and includes a scripting language for the batch processing of speech files. Netlab is a neural network toolbox that supports a wide range of data analysis techniques. It is used in conjunction with Matlab, a mathematical modelling language [11]. Neural networks provide an effective means of modelling complex relationships between data sources [12]. An important property of neural nets is that their performance is highly resistant to the effect of ‘noisy’ data (i.e. input parameters that contain spurious values).

For a neural network to ‘learn’ how to estimate gender and age, it must be trained using a data set of speakers whose gender and age are known. In our case, we used a subset of the CTIMIT database [13]. This speech database contains single-sentence utterances recorded in a mobile setting. We used only the recordings made by speakers in the age range 21–55, resulting in 3303 recordings spread across 621 speakers. Previous studies did not make use of mobile phone recordings [6, 7].

3.2 Procedure and Evaluation

A Praat script was used to automatically extract estimates of the relevant acoustic features from the recordings. Thirteen measures were extracted for each recording: five values for jitter, six for shimmer, the mean fundamental frequency (F_0) and the mean harmonics-to-noise ratio (HNR). These values were combined with the known gender and age of each speaker to form a 15 element vector for training the neural network. The network was trained on 80% of the recordings, validated against a further 10% and evaluated using the remaining 10%. We experimented with a number of network configurations; the learning parameters of the best performing network were: 20 hidden layer units, learning rate=0.2, momentum=0.2, training iterations=200).

With respect to gender, the estimator program performed very well. It correctly classified 94.4% of the test cases. In comparison with a simple classifier that always predicted the most frequently occurring value, male (69.1%), the estimator still performed significantly better (two-tailed t-test, $p < 0.01$). Looking at age prediction, the mean error of the test cases was -0.1 but the standard deviation was 6.86. In other words, the neural network failed to learn a useful relationship between the acoustic features and speaker age. We have identified three possible reasons for this result. Firstly, there simply may not be a significant variation in the acoustic features within the age range 21-55. Secondly, the acoustic features that were used may be excessively influenced by background noise. Thirdly, the distribution of speaker age within the CTIMIT database is skewed towards speakers in their 20s and 30s. More training examples of speakers in their 40s and 50s may be required.

4 Summary and Further Work

In this paper we examined a number of acoustic features for profiling mobile users of conversational interfaces. Specifically, we investigated whether a user's gender and age could be estimated in spite of a high level of background noise. We implemented an automatic estimator using acoustic feature extraction and neural network applications. We tested the program using recordings of mobile phone users. The estimator program achieved a very high level of performance with respect to gender but failed to estimate age to a significant level of accuracy.

In the near future, we hope to collect more examples of speakers over the age of 40. This should provide a clearer assessment of the potential for estimating speaker age. We will also investigate other characteristics for profiling users of mobile devices. Potentially valuable user traits include physical size, emotional state, rate of speaking and identification of the user's native language. In the longer term, we intend to integrate these results into a single user profiling module and make it available to developers of conversational interfaces.

References

1. Zue V. and Glass J. Conversational Interfaces: Advances and Challenges. *Proceedings of the IEEE* **88**(8), 2000. pp 90–169.
2. Marmasse N. and Schmandt C. Location-aware information delivery. In *Proceedings of IEEE International Symposium on Handheld and Ubiquitous Computing, Bristol, UK. September 2000*.
3. Arnold K. and Bianchi C. Relationship Marketing, Gender, and Culture: Implications for Consumer Behaviour. *Advances in Consumer Research*, **28**, 2001. pp. 100–105.
4. Eskenazi M. and Black A. A study on speech over the telephone and aging. In *Proceedings of Europeech 2001, Aalborg, Denmark. September 2001*.
5. Gong Y. (1995): Speech Recognition in Noisy Environments: A Survey. *Speech Communication* **16**(3), 1995. pp. 261–291.
6. Minematsu N., Sekiguchi M. and Hirose K. Automatic Estimation of Perceptual Age Using Speaker Modeling Techniques. In *Proceedings of Europeech 2003, Geneva, Switzerland. September 2003*.
7. Müller C., Wittig F. and Baus J. Exploring Speech for Recognizing Elderly Users to Respond to their Special Needs. In *Proceedings of Europeech 2003, Geneva, Switzerland. September 2003*.
8. Ferrand C. Harmonics-to-Noise Ratio: An Index of Vocal Aging. *Journal of Voice*, **16**(4), 2002. pp. 480–487.
9. Praat, <http://www.praat.org>.
10. Netlab, <http://www.ncrg.aston.ac.uk/netlab/>.
11. Matlab, <http://www.mathworks.com/products/matlab/>.
12. Bishop C. *Neural Networks for Pattern Recognition*. Oxford University Press, Oxford, 1995.
13. Brown K. and George E. CTIMIT: A speech corpus for the cellular environment with applications to automatic speech recognition. In *Proceedings of ICASSP 1995, Detroit, US. May 1995*.