

USER VOICE IDENTIFICATION (URVIN)[§]

AMADEUS VIRTUAL RESEARCH CENTRE

Aladdin Ariyaeinia, Reza Sotudeh, Chris Bailey* and Fredrik Rodin**

University of Hertfordshire, *University of York, **Fulcrum Voice Technologies

INTRODUCTION

A factor important to the success of a variety of ubiquitous computing applications is that of identification of users so that each individual is provided with an appropriate level of authorisation. This can be for a variety of purposes both at home and work environments. Examples are controlling appliances, physical access to restricted areas, access to electronic information, launching software programmes, access to data over the Internet and online financial transactions.

A difficulty with the conventional means of identification such as passwords, personal identification numbers and ID cards is that they are not designed (therefore are not suitable) for use in smart environments. Secondly, the adaptation of facilities to operate with these means of identification may result in such front-end subsystems which require complex human interaction. Moreover, such identification means can be easily compromised. In view of these, it appears that the required optimal usability and reliability in determining the identities of users may only be achieved through the deployment of user-friendly biometrics. An identification method in this category is speaker recognition (voiceprint).

A main component of any speaker recognition system, and also any speech-based interactive system in general, is the speech feature extraction engine (FEE). For the purpose of practical applications, however, such a feature extraction engine should incorporate effective means for the reduction of the effects of variations in speech characteristics. This is mainly to enhance the speaker identification performance in uncontrolled environments where background noise can significantly reduce the identification accuracy.

AIM AND APPROACH

The purpose of this project is to investigate an effective hardware system capable of extracting highly reliable speech features in the presence of background noise. It is envisaged that the development of such a reusable hardware IP-Core will have many benefits for ubiquitous computing applications, including rapid development and reduction in cost vs. functionality for future products incorporating voice-driven interactions.

The hardware implementation will use the system on chip (SoC) approach, based on a standard FPGA/DSP platform. This will allow the greatest flexibility in the choice of programmed versus application-specific implementation, allowing all architectural options to be explored. The SoC approach should also leave sufficient overhead, for instance to allow fusion of voice with image or video recognition functions. The AMADEUS Centre (Architectures Machines And Devices for Efficient Ubiquitous Systems), manages a programme of work to develop a number of such re-usable IP-blocks for future Ubiquitous systems development. This includes data and pattern matching Ip-Cores, Video Feature Extraction, and networking support functions, among others. Integration of Audio Feature-extraction with CPU and/or a pattern-matching engine could yield highly integrated voice identification solutions for future application in the pervasive computing domain. Work in AMADEUS includes development of ubiquitous system development platforms, which will make use of low-footprint CPU technology supplied by silicon partners.

CHALLENGES IN AUTOMATIC SPEAKER IDENTIFICATION

The process of automatic speaker recognition can be defined as the extraction of the personal identity information from a presented sample utterance using signal measurement techniques. Two sub-classes of speaker recognition are speaker verification and speaker identification. The former is to determine whether a speaker is who he or she claims to be. Speaker identification is the process of determining the correct speaker from a given population. Each of the above two sub-classes can be either text-dependent or text-independent. In the former mode, the user must provide utterances of the same linguistic content for both training and recognition. In the latter mode, however, speakers are not constrained to provide utterances of specific texts for recognition.[§]

The current project is concerned with *open-set, text-independent speaker identification* (OSTI-SI). Given a set of registered speakers and a sample utterance, open-set speaker identification is defined as a twofold problem. Firstly, it is required to identify the speaker model in the set, which best matches the test utterance. Secondly, it must be determined whether the test utterance has actually been produced by the speaker associated with the best-matched model, or by some unknown speaker outside the registered set. The difficulty in this problem is exacerbated if the process is text-independent.

The problem of OSTI-SI is further complicated by undesired variations in speech characteristics due to anomalous events. These anomalies can have different forms ranging from the environmental noise to uncharacteristic sounds

[§] Further information on URVIN can be found at: <http://www.cs.york.ac.uk/amadeus/>

generated by the speaker. The resultant variations in speech cause a mismatch between the corresponding test and pre-stored voice patterns. This can in turn lead to degradation of the OSTI-SI performance. The potential errors and difficulties in OSTI-SI can be analysed as follows.

Suppose that N speakers are enrolled in the system and their statistical model descriptions are $\lambda_1, \lambda_2, \dots, \lambda_N$. If \mathbf{O} denotes the feature vector sequence extracted from the test utterance, then the open-set identification can be stated as:

$$\max_{1 \leq n \leq N} \{p(\mathbf{O} | \lambda_n)\} \geq \theta \rightarrow \mathbf{O} \in \begin{cases} \lambda_i, i = \arg \max_{1 \leq n \leq N} \{p(\mathbf{O} | \lambda_n)\} \\ \text{unknown speaker model} \end{cases}, \quad (1)$$

where θ is a pre-determined threshold. In other words, \mathbf{O} is assigned to the speaker model that yields the maximum likelihood over all other speaker models in the system, if this maximum likelihood score itself is greater than the threshold θ . Otherwise, it is declared as originated from an unknown speaker. It is evident from the above description that, for a given θ , three types of error are possible:

- \mathbf{O} , which belongs to λ_m , not yielding the maximum likelihood for λ_m .
- Assigning \mathbf{O} to one of the speaker models in the system when it does not belong to any of them.
- Declaring \mathbf{O} which belongs to λ_m , and yields the maximum likelihood for it, as originated from an unknown speaker.

These types of error are referred to as *OSIE*, *OSI-FA* and *OSI-FR* respectively (where *OSI*, *E*, *FA* and *FR* stand for open-set identification, error, false acceptance and false rejection respectively). Based on equation (1), it is evident that open-set identification is a two-stage process. For a given \mathbf{O} , the first stage determines the speaker model that yields the maximum likelihood, and the second stage makes the decision to assign \mathbf{O} to the speaker model determined in the first stage or to declare it as originated from an unknown speaker. Of course, the first stage is responsible for generating *OSIE* whereas, both *OSI-FA* and *OSI-FR* are the consequences of the decision made in the second stage.

An important point to note in this two-stage process is that the latter stage is far more susceptible to distortions in the characteristics of the test utterance than is the former stage. This is because, in the former stage, since the same test utterance is used to compute all the likelihood scores, the distortions in the test utterance are likely to be similarly reflected in all the likelihood scores. As a consequence, the selection of the model that yields the maximum likelihood is likely to be unaffected. On the other hand, in the second stage, the absolute maximum likelihood score is compared against a threshold determined a priori and without any knowledge about the characteristics of the distortion in the test utterance.

It should be pointed out that a task similar to that described above (in the second stage of open-set identification) is also encountered in speaker verification. However, in this case, the problem is not as challenging. To be more specific, the challenge in open-set identification can be viewed as a special (but unlikely) scenario in speaker verification in which each impostor targets the speaker model in the system for which he/she can achieve the highest score. This point is further illustrated by Figure 1 which shows typical score distributions associated with these two forms of speaker recognition under the same experimental condition. As observed, the overlapping between the score distributions for unknown and known speakers in open-set identification is considerably greater than that between the score distributions for impostors and true speakers in speaker verification.

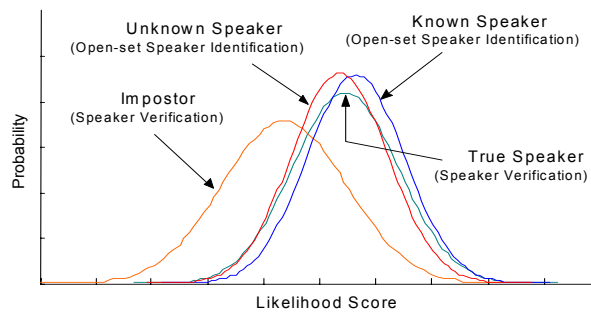


Figure1: Score distributions associated with speaker verification and the second stage of open-set speaker identification.

VISION STATEMENT

The work is currently focused on the development of algorithms for the extraction of reliable speech features for speaker identification. The outcome of this work is expected to include effective statistical methods for speech enhancement and also the feature extraction techniques which help maximise discrimination amongst speakers. A feature extraction engine will subsequently be developed by translating the algorithms into a hardware architecture. It is envisaged that the development of such a feature extraction engine, with careful reference to the requirements in speech processing applications, will lead to a generic IP core. The opportunity to integrate this with other SOC components into complete systems then becomes achievable. In the ubiquitous systems domain, achieving high function density with low-footprint hardware solutions is an essential goal, which we believe will be aided by our work.