# Position Paper on Speaker Biometrics and VoiceXML 3.0

Qiru Zhou
Bell Laboratories, Alcatel-Lucent
600 Mountain Avenue
Murray Hill, NJ 07974
USA
qzhou@research.bell-labs.com

## INTRODUCTION

Speaker identification (SI) and speaker verification (SV) are integral parts of natural voice communication functions. We believe that VoiceXML version 3.0 must include SI and SV functions for web service applications with voice dialog. Based on expertise and experience in speaker identification and speaker verification research in Bell Labs, we are interested in participant W3C voice browsing workgroup effort to work on speaker biometric standard to promote these technologies to real-world applications.

## TECHNICAL CLARIFICATION

Although there are similarities between SI and SV, we want to clarify the differences between them that we believe need different treatment in a markup language:

SI is the process of associating an unknown speaker with a member in a population; i.e., a multiple-choice classification problem.

One example of SI application is to identify an active speaker in an audio conference that all participants are known to the conference.

SV is the process of verifying whether an unknown speaker is the person as claimed; i.e., a yes-no hypothesis testing problem.

Example of an SV application is using voice as password to access a restricted system.

## SI AND SV TECHNOLOGIES

There are several pattern matching technologies may be used for SI and SV, in Bell Labs research, we believe that Hidden Markov Model (HMM) based statistical modeling method is an effective, high accurate and flexible method to use for SI and SV to support fixed phrase, prompt phrase and text independent SI and SV. We are interested in to see VoiceXML 3.0 specification have sufficient support of HMM based SI and SV.

INTEGRATION OF XML AND NON XML FORMATS

a. Should speaker biometric data be represented in binary or xml formats? Should both formats be standardized?
   There are two types of biometric data: the speaker voice sample and speaker voice print model data for pattern matching. The former can be easily standardized (encryption is a must for security), the latter involves standardize detailed SI and/or SV algorithm that may not be practical to support multi-vendors don't share their SI and/or SV engines. BioAPI treats it as "binary blob" that leave vendor to define this part. For efficiency, binary should be supported.
b. Should developers use a programming API or an XML-based API? Should both formats be standardized?
   A programming API is suitable for stand alone and certain client-server based applications. For web services based applications, we need XML based API. I am not sure if a programming API is in the scope of VoiceXML.
c. Can commands, events, and data structures be normalized or standardized?
   We believe it is feasible to standardize these.
d. What is the relationship with MRCP V2 and how should that relationship evolve?
   As MRCPv2 draft states "MRCPv2 enables the implementation of distributed Interactive Voice Response platforms using VoiceXML browsers or other client applications while maintaining separate back-end speech processing capabilities on specialized speech processing servers." VoiceXML may work together with MRCP V2, or they can work independently to implement applications. Coordination effort is required to support the first scenario.
e. What is the relationship with BioAPI and how should that relationship evolve?
   BioAPI is in the category of a programming API. It can be used as a server side API, or stand alone application API. In order to integrate BioAPI and VoiceXML 3.0 SI and SV functions, we need a mapping relation between them.
f. Can audio formats be normalized for interoperability and conformance testing? How to determine that normalization?
   We believe this question is yes. Waveform based audio is straight forward and feature vector based audio samples such as ETSI DSR format may be used for this purpose.

SECURITY DATA

a. What mechanisms should be used to maintain the security of voice models and voice model databases?
b. What mechanisms should be used to maintain the additional non-biometric security data?
c. What mechanisms should be used to secure the transferring of voice biometric data between client and server?

In general, HMM based SI/SV voice models is a one way transformation that is impossible to recover back to the original voice samples. We believe IP and web based encryptions are sufficient if model need to transport on network, including biometric and non-biometric portions of the speaker voice model data.

## THE ROLE OF MULTIMODAL BIOMETRICS

a. How to combine biometric and/or nonbiometric techniques to reach authentication decisions?
b. What is the role of multimodal technology in authentication?
c. What is the role of EMMA or other multimodal annotation techniques?

Due to imperfectness of today's SV technology, we believe voice based dialog or multimodal dialog (especially for mobile applications) are required to support failover of SV authentication.

## CONCLUSION

We believe that SI/SV functions are essential for a natural voice user interface specification. We are interested in work with internet research and engineering community to contribute to W3C VoiceXML standard.
Due to current resource constraint, we would like to contribute at SI/SV research expert level at present.

## DISCLAIMER

This paper is an expert view of speaker identification and speaker verification for W3C workshop discussion purpose. It does not represent the official position of Alcatel-Lucent.