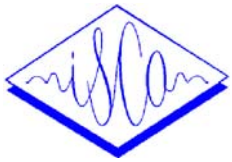

ODYSSEY04 - The Speaker and Language Recognition Workshop
Toledo, Spain
May 31 - June 3, 2004

"Text-prompted" Without Text: a Language-independent Voice-prompted Speaker Recognition System

Yaakov Metzger, Ran Gazit

Persay Ltd.
22a Raul Wallenberg St., Tel Aviv 69719, Israel
Ran_Gazit@persay.com

Abstract

A new paradigm of voice prompted speaker recognition is presented. The vocal prompts that the speaker is asked to repeat are used by the speaker recognition system for segmenting the data and for normalizing the verification results. Using the vocal prompts themselves instead of the matching text makes the overall system more flexible and truly language independent. A technology demonstration system was set up and a small-scale experiment measured both speaker verification and text confirmation performance. Testing results show good performance when using human vocal prompts as well as synthesized vocal prompts.

1. Introduction

Speaker Recognition systems are often categorized according to the restrictions on the text used for enrollment and verification. In a text-independent speaker recognition system, there are no such restrictions, and speakers are free to use any text they wish. Text independent speaker recognition systems are therefore used when speakers are engaged in free speech conversation, such as the conversation between a customer and her bank agent. In most cases, text-independent speaker recognition systems do not require labeling or segmenting the speech data into words or phonemes. Most text-independent systems are therefore also language independent.

The situation is different in text dependent systems. In this case, speakers are asked to repeat a specific text known to the system. Knowledge of the text can be exploited to improve performance [1]. The text may be used during both enrollment and verification for segmenting the incoming speech to smaller units (words, sub-words or phonemes). During verification, the system may also verify that the speaker actually said the required text (text confirmation), in addition to verifying the claimed identity of the speaker (speaker verification). Text dependent speaker recognition systems are used when the speaker is communicating with a computer or other intelligent yet artificial device, such as an Interactive Voice Response (IVR) system. Text dependent systems that require labeling or segmentation of the data are language dependent, as the segmentation is usually done according to a language-specific model.

In a standard text dependent system, the text used for enrollment and for all verification trials is identical. A malicious impostor may use a recorded conversation between a speaker and the system, and play back the recorded speech in order to gain the privileges of the recorded speaker. To reduce the possibility of playback attacks, a text-prompted approach is used [2,3]. As in the text dependent system described above, the speaker is still asked to repeat a text known to the system. However, this text is different each time the user calls the system. The system accepts the claimed identity of the speaker only when it decides that the speaker has uttered the prompted text. This method recognizes the speaker, but should also reject utterances whose text differs from the prompted text.

Most text-prompted systems use strings of single-digit or two-digit numbers, and generate a new random string each time the speaker calls the system. During the enrollment phase, utterances are cut into smaller units, such as words (or digits) [3-6], sub-words [7,8], tri-phones [9] or single phonemes [2,10-13]. All the occurrences of each unit in the enrollment data are grouped, and a speaker-dependent model is trained for each unit. In the recognition stage, a new text based on a different sequence of the same units is formed, and the user is asked to repeat it. The system concatenates the unit models of each speaker according to the prompted text, and a speaker-dependent sentence model is created. The likelihood of the input speech against that sentence model is calculated and used for speaker recognition. This likelihood is often normalized by the likelihood of the input speech against a speaker-independent (universal) sentence model. This background model is built from concatenation of speaker-independent models of each unit according to the same prompted text.

Segmentation of the enrollment data or background data into units, as well as concatenation of units according to text is based on a set of rules which are, in most cases, language specific. Even when using a decoupled combination of speaker and speech recognizers [9,13] to achieve the dual goal of text confirmation and speaker recognition, the speech recognizer is necessarily language dependent. In many operational scenarios, language dependency and a pre-defined closed vocabulary are constraints that must be relieved.

This paper suggests a flexible, language-independent form of text-prompted speaker recognition, without using the text itself (i.e. without using the phonetic alphabet description of the text or similar representation). Instead, the system uses the acoustic representation of the vocal prompts that the speaker is asked to repeat. These prompts are prepared by the system operator prior to opening the service, either by recording a professional annotator, or by using a speech synthesizer. The next section shows how these vocal prompts may be used in a "voice-prompted" speaker recognition system instead of the underlying text. Since phonetic alphabet and similar

representations are language dependent, this approach relieves the language dependency constraint and the only adaptation to a new language consists of recording new prompts.

2. System Description

2.1. Concept

Before installing a text-prompted speaker verification system, the operator prepares a set of vocal prompts that will instruct the users as to what should they say. For example, the operator may prepare a separate vocal prompt for each English digit. During enrollment or verification the system will select a random sequence of digits, and the IVR will concatenate the vocal prompts of the matching digits and play them to the user. The user will be asked to repeat these vocal prompts in her own voice.

The vocal prompts can be recorded by a professional annotator or synthesized by a Text-to-Speech (TTS) system. In a standard text-prompted system, the lexicon from which the individual tokens are selected is limited in size and language, and is the same for all speakers.

Usually, the vocal prompts are used only for directing the speakers as to what should they say. However, these vocal prompts carry with them information that can be exploited by the speaker recognition system. Specifically, acoustic models built directly from the vocal prompts can be used for initial segmentation of the incoming data, and also as speaker independent models for score normalization.

This concept actually tests if the speaker had repeated exactly the same prompts played to her, without looking at the text these prompts represent. This is a prompted speaker recognition system, but it is not text-prompted, as the actual text is not known to the system. We suggest to denote this concept as "voice-prompted" speaker recognition.

This concept allows arbitrary selection of tokens, with no restriction on the number of possible tokens, the lexicon, or the language. Furthermore, different sets of tokens may be easily used for different speakers, as long as the vocal prompts played to the speaker are also fed to the speaker verification system.

2.2. System description

The system is using a large set of words that were prerecorded by few human annotators, or synthesized by a TTS engine using several different "voice types". The word set is not limited, and can contain any type of words in any language. Each word is recorded by all annotators (human or synthesized), and a Hidden Markov Model (HMM) is generated for each word from all the recordings of that word. One set of recordings will be used as the vocal prompts played to the users.

When a user enrolls to the system, a small random sub-set of words is chosen. This is the user-specific vocabulary, and it can be different for each user or group of users. The user will be prompted to repeat a few sequences of the words from her vocabulary. The sequences should include each one of the words in the user-specific vocabulary at least three times, each time in a different context. Instructing the user to repeat the sequences is done by concatenating the vocal prompts of the matching words, and playing them to the speaker.

A speaker-independent model is built for each enrollment sequence by concatenating the HMM's built from the annotators' vocal prompts. The speaker-independent sequence model is used to segment the speaker's data and label the voice segments that match each of the words in the sequence. When enrollment recordings are done, all the repetitions of each word are collected, and a speaker-specific HMM is trained for each word. These HMM's, together with the indices of the words that form the speaker-specific vocabulary will be saved in the database of the system as the voice signature of each speaker.

When a verification call is initiated, the system downloads the voice signature of the speaker from the database, and reads the speaker's HMM's, as well as the indices of the words in the speaker's vocabulary. A short, random sequence of a sub-set of these words is generated, and the speaker is asked to repeat a sequence of the matching vocal prompts. A speaker-specific HMM is built for that sequence by concatenating the matching speaker's HMM's. In a similar way, a speaker-independent "background" model is built by concatenating the matching speaker-independent HMM's. The utterance is verified by computing the likelihood against each of the models, and using the log-likelihood ratio as a final score that should be compared with a threshold for decision. For additional security, this process may be repeated with another sequence of words drawn from this speaker's vocabulary.

2.3. Call flow

The following dialog illustrates a possible enrollment scenario. The token set in this example is based on the NATO Phonetic Alphabet (used in radio communications when sending information that needs to be spelled). The first five letters were chosen as the vocabulary of this specific speaker, and the speaker is asked to repeat each letter three times, each time in a different context:

System: "Please type your account number".
Caller: 1234 (input by DTMF signaling)
System: "Please repeat: echo alpha bravo".
Caller: "*echo alpha bravo"*
System: "Please repeat: alpha delta charlie".
Caller: "*alpha delta charlie*".
System: "Please repeat: "delta bravo echo".
Caller: "*delta bravo echo*".
System: "Please repeat: echo charlie delta".
Caller: "*echo charlie delta*".
System: "Please repeat: alpha charlie bravo".
Caller: "*alpha charlie bravo*".
System: "Thank you, you had successfully completed the enrollment process".

When this speaker, at a later stage, calls the system for verification, the following schematic dialog might take place:

System: "Please type your account number ".
Caller: 1234 (input by DTMF signaling).

System: "Please repeat: "bravo charlie delta".
Caller: "*bravo charlie delta*".
System: "Thank you, your voice was verified", or:
"Sorry, your voice was not verified."

It should be noted that each time the speaker calls the system for verification, the system asks for a random sequence of three out of the five possible words. A different speaker enrolled to the system might be trained with a different set of five words, out of the twenty-six possible tokens that were recorded in this case into the system. The numbers five and three were used here as an example. The system can be set to choose a larger set of words for each speaker, and/or ask the speaker to repeat longer sequences. The limitations here are:

- The longer the sequence, the more challenging it is for the speaker to repeat.
- A larger set of words would require a longer enrollment session, since at least three enrollment repetitions of each word are required to train its HMM.

The following section describes a small scale experimental evaluation of this concept.

3. Experiment setup

In order to test a voice-prompted system, we need a database with recordings of speakers that respond to actual vocal prompts. We assume that the prompts have some effect on the way speakers pronounce their utterances. There is an unconscious tendency to imitate the pronunciation and tempo of the annotator, and therefore actual prompted collection should be done, while reading text from paper is not adequate.

We had set up an audio collection system that used the same vocabulary of five English words for all speakers (our creative minds chose the words 'one', 'two', 'three', 'four' and 'five'). This way we could later test each speaker as an impostor to all other same-gender speakers. Each speaker called the system five times. Each call followed the enrollment protocol, and prompted the speaker to repeat five different sequences of three words. Each word appeared exactly three times in each call.

The first call of each speaker was used for enrollment and the 4 remaining calls were separated into individual sequences that were used for verification experiments. Segmentation of the calls to words, for both enrollment and verification trials, was done with HMM models built from the vocal prompts.

All calls were collected through land-line telephones. Audio preprocessing in this evaluation system includes 8KHz sampling, and calculating 20 LPC CEPSTRA features on 25 ms frames overlapping by 12.5ms. All feature vectors of an utterance are normalized by subtracting the average feature vector and dividing by the standard deviation vector. The model for each word is a left-to-right HMM, initialized with uniform segmentation.

The verification tests were separated into 4 groups: correct response of the target speaker, which the system is supposed to accept, and three attack scenarios which the system is supposed to reject:

- The first attack scenario is an impostor repeating the correct prompt. To simulate this attack we verified each test utterance against models of other speakers that were concatenated according to the spoken utterance.
- A second attack involves the correct target saying a different prompt. Rejecting this kind of attack is the advantage of a prompted system, as it makes it immune to playback attacks. We simulate this attack by verifying each test utterance against a random sequence of the same speaker's models.
- The third attack is a trivial attack of an impostor that does not even say the correct prompt. This should be easily rejected, but should be tested, because the wrong sequence will affect the "background" model score as well, and therefore might improve the final score. This is simulated by verifying each test utterance against random sequences of models of other speakers.

The analysis was repeated four times using different sets of vocal prompts. The vocal prompts are used to create speaker-independent "background" models, which are used for segmentation of the enrollment and verification utterances, and for normalizing the verification score. The first set of vocal prompts was based on voices of two male and two female human annotators, obtained with desktop quality recording (not through a telephony network). The next set of vocal prompts was created by a formant-synthesis TTS system, using two male voices and two female voices. This was repeated with a concatenative TTS system. We also used the concatenative TTS with vocal prompts of only one male speaker and one female speaker. Formant synthesis generates highly intelligible, but not completely natural sounding speech, while concatenative synthesis has higher potential for sounding "natural" [14,15].

Table 1 presents the effect of various sets of vocal prompts on the equal error rate (EER), for the first attack scenario (impostor is repeating the correct utterance). We assume that the system is tuned to work at the equal error point for this attack scenario, and compute the false accept rate at this working point for the other two attack scenarios (target speaker and impostor repeating incorrect password). This false accept rate is shown in Table 1 for all sets of vocal prompts.

	Correct text	Wrong text	
	EER	False accept	
		Target	Impostor
2M+2F, human	9.3%	9.3%	0.2%
2M+2F, f_TTS	7.4%	4.9%	0.1%
2M+2F, c_TTS	6.5%	4.9%	0.3%
1M+1F, c_TTS	7.1%	2.2%	0.1%

Table 1: Error rates. (c_TTS- concatenative Text-to-Speech, f_TTS- formant synthesis Text-to-Speech).

Figure 1 shows three Detection Error Tradeoff (DET) curves for the 2M+2F, concatenated TTS case. The curves show the miss probability, which is the probability that a target speaker saying the correct text is rejected, against the false alarm probability. For each threshold value, there are three possible values of false alarm, according to the three attack scenarios described above. The system is required to reject all three attack scenarios, using a single decision threshold, while accepting correct target response. Each threshold determines the "miss probability", and the resulting "false accept probability" appears on the intersection of the three DET curves with a horizontal line that passes through the EER point of the upper most curve.

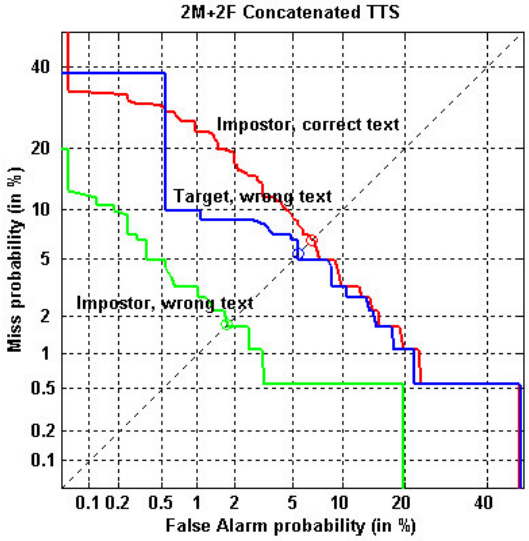

Figure 1: DET curves for the 2M+2F concatenative TTS case, under three possible attack scenarios.

This analysis shows minor performance differences between the various sets of vocal prompts. Considering the size of the database and the resulting accuracy all sets obtained similar EER in the major task of speaker verification (both target speaker and impostor uttering the correct text), with slight advantage to the concatenative TTS with multiple voice types. Note that this EER value is based on speaker verification over a single sequence, only three-words long. The inferiority of the human recorded prompts to TTS prompts is not fully understood. A possible explanation is that the different voices of the TTS system overlap better in their temporal structure, and thus produce more accurate models.

Text confirmation performance (target speaker uttering the wrong text) is similar to the speaker verification performance, with some advantage to the concatenative TTS. Note that this text confirmation ability was reached without using the text itself, and with no language model or other additional information. Finally, rejecting casual impostors uttering the wrong text seems trivial with this system.

4. Conclusion

A voice-prompted alternative to the text-prompted speaker verification concept is suggested, in which vocal prompts are used for audio segmentation and score normalization. A limited performance evaluation showed good performance, for both speaker verification and text confirmation tasks. This concept allows language-independent prompted speaker recognition, with no constraints on the lexicon size or type. This flexibility allows using various sets of passwords for different groups of speakers, for maximum personalization of the speaker recognition service.

Future studies of this concept will compare its performance to the standard text-prompted approach over a common database. Additionally, large-scale tests should be conducted to measure the effectiveness of this concept under various channel conditions and with various sets of vocal prompts.

5. References

[1] G.R. Doddington, "Speaker Recognition – Identifying People by their voices", Proc. Of the IEEE, Vol. 73. No. 11. Nov. 1985, pp. 1651-1664

[2] A. Higgins, L. Bahler and J. Porter, "Speaker Verification Using Randomized Phrase Prompting", Digital Signal Processing, Vol. 1., pp. 89-106, 1991

[3] S. Furui, "An Overview of Speaker Recognition Technology", ESCA Workshop on Automatic Speaker Recognition, Identification and Verification, 1995

[4] A.E. Rosenberg et. al., "Connected Word Talker Verification Using Whole Word Hidden Markov Models", ICASSP 1991

[5] J. Lindberg and H. Melin , "Text-prompted versus Sound-prompted Passwords in Speaker Verification Systems", Eurospeech 97, Vol. 2, pp. 851 – 854

[6] T. Masuko et. al., "On the Security of HMM-Based Speaker Verification Systems Against Imposture Using Synthetic Speech", Eurospeech 99, Vol. 3, pp. 1223-1226

[7] D. James et. al., "CAVE – Speaker Verification in Banking and Telecommunications"

[8] W.M. Campbell and C.C. Broun, "Text-Prompted Speaker Recognition with Polynomial Classifiers", Odyssey 2001

[9] D. Reynolds and B.A. Carlson, "Text-Dependent Speaker Verification Using Decoupled and Integrated Speaker and Speech Recognizers", Eurospeech 95, pp. 647-650

[10] T. Matsui and S. Furui, "Concatenated Phoneme Models for Text-Variable Speaker Recognition", ICASSP 1993

[11] T. Matsui and S. Furui, "Speaker Adaptation of Tied-Mixture-Based Phoneme Models for Text-Prompted Speaker Recognition", ICASSP 1994

[12] C. Che et. al., "An HMM Approach to Text-Prompted Speaker Verification", ICASSP 1996

[13] A. Lodi et. al., "Very Low Complexity Prompted Speaker Verification System Based on HMM Modeling", ICASSP 2002

[14] T. Dutoit, "A Short Introduction to Text-to-Speech Synthesis". TTS research team, TCTS Lab., Belgium, 1996. http://tcts.fpms.ac.be/synthesis/introtts.html

[15] J. Schroeter, "The Fundamentals of Text-to-Speech Synthesis", VoiceXML review, Vol. 1, No. 3, Mar. 2001