The Importance of One’s Speech Recognition in Electronic Services

Azamolsadat Parei
Information Technology Engineering Group, Department of Industrial Engineering, K. N. Toosi University of Technology
as.parei@kntu.ac.ir

Hojjatollah Hamidi
Information Technology Engineering Group, Department of Industrial Engineering, K. N. Toosi University of Technology
h_hamidi@kntu.ac.ir

Abstract

Regarding the development obtained in speech recognition and processing sciences, this application can be used in electronic services. One of the fields within which biometrics are applied is to use one’s speech that has been noted more during last decade. The fingerprint, eye and speech are biometrics with usual applications amongst others. Regarding the fact that the speech can be used remotely, it is possible to use it in remote electronic services, too. Hence, it is considered as an efficient method in different telephone applications in particular electronic banking systems. Therefore, it may be used for credit authentication when fulfilling banking transactions, checking the account balance and etc. through cellphone. However, using speech recognition and speaker authentication have faced several challenges including sound emulation by forgers, recorded sounds, sensitivity to various noises and need to large amount of data for test. This research aims at the importance of speech recognition to be used in mobile and electronic banking.

Keywords: electronic service, speech recognition, mobile banking, non-speaker data, speaker authentication.
Introduction

Technological development during recent years has been resulted in profound changes over various technologies one of which is people recognition using their biometrics. Biometry is the technique of recognizing people using their physical and behavioral features. Security requirements have made biometry to grow faster recently. More prevalent method of biometry includes recognition by way of finger print, face, iris, retina and speech. Among those methods, finger print is the most applied one. However, using people’s voice print has extremely been noted because it is the sole method that can be used remotely. Using speech for remote authentication has been facilitated by increasing cellphones and as a result made this method to be universal in particular for mobile banking. Using the voice print for authentication is performed in two steps: registration and registration verification or speaker’s identity. Registration is an active process within which a speech sample is taken for the purpose of making a biometric pattern out of user’s identity. The pattern is consequently saved and used for verification by way of comparison in next contacts (Nancie Gunson, Diarmid Marshall, Fergus McInnes, Mervyn Jack, 2011).

Speaker authentication

As everybody knows, the speaker authentication refers to a process through which one’s claim is accepted or rejected by his/her voice. This is performed text-dependent or text-independent. In text-dependent mode the registration and test are performed simultaneously while in other mode there is no limit imposed on the content. The dependent method is more accurate than the independent one and is usually used foe security applications. The sound emulation by forgers is one of the most challenging problems. In case that the forger’s voice is much more similar to the target’s identifying counterfeit will be a difficult issue. Other challenges include recorded sounds for later misuses. Normally, a process is included into the authentication system to prevent sound record. The tsrk of the process is to recognize the recorded sound. There is still another problem for speaker authentication. That is the effect of transmission channel on the people’s speech. Those effects change spoken characteristics and as a result deteriorate the accuracy of those systems. On more challenge is that a long speech is needed to reach a reasonable accuracy.

Speaker recognition classifications

Speaker recognition definition

Speaker recognition refers to a process that a person is recognized using his/her voice and the claim is ascertained. In this process, one’s voice is recorded as the training voice during training time. The, properties of the voice including intonation, sound, pitch are extracted. After that, a model is devised for anybody. At the time of test, the entering voice is taken and those properties are extracted from. The, they are scored based on those models. Those speakers with higher score will be selected as the target speaker. In order to use these systems an accuracy exceeding 89% is required. If this accuracy isn’t considered when authentication is performed it is likely that the users’ account number or their account code to be hacked. Speaker recognition has two sub-groups: speaker identification and speaker authentication.
Speaker identification
There are a number of speakers who have been registered on the system. When testing the voice for an unknown speaker is received. Now, using this voice it must be determined that to whom this voice is belonged.

Speaker identification in turn is performed in two ways:
Closed-set: in this method, the entering speech style is belonged certainly to a speaker who has been registered on the system. This method is easier than the other one. It is suffice here to obtain the score of the entering speech for all speakers’ models. Here, those speakers with higher scores will be determined as the target.

Open-set: It is possible that the entering voice no to be belonged to any speaker registered on the system. This is more difficult comparing to the closed-set. Just like the closed-set in this method, scores of all speakers are also obtained. There are two thresholds in these systems obtained at training time. At that time, the highest score is first obtained as well as the speaker with highest score. If the score exceeds the speaker’s highest score he/she will be recognized as the target. If the speaker’s score is lower than the first threshold but higher than the second one he/she will be recognized as doubtful speaker. However, if the score is less than the second threshold it will be concluded that this speech belongs to no one of speakers on the system (Nancie Gunson, Diarmid Marshall, Fergus McInnes, Mervyn Jack, 2011)

Speaker identification is harder than the authentication. However, except for minor problems including recording people voice stealthy there are several methods reducing the risk of this attack. Those methods are essentially based on live voice which include signal processing techniques to analyze speech properties as well as recognize the recorded voice. Since certain factors in speech are innate for people one of the suitable methods is to compare current authentication with previous suggestions which must be down at the registration phase in addition to comparing created voice ID. In addition, requesting random numbers used when registering which are used at the time of approval strengthen the security. Therefore, there are considerable resources preventing fraudulent. Thus, in case of mobile banking within which the accuracy is essentially important a text-dependent strategy performs well compared to text-independent one and is among fraudulent preventing methods (Zin Ali, 2004).

Speaker authentication
Authentication is a quite different issue. Here, one primarily claims that he is one of system speakers like Amir. After that he is asked to speak little. Then the score of recorded speech is computed by Amir’s speech model and compared to Amir’s first threshold. If exceeds it the system approves his claim regarding being Amir otherwise rejects it. Speaker authentication is classified into two subclasses: text-dependent and text-independent.

Text-dependent: in this method, the speaker must repeat a known statement at the time of training and the system must proclaim the result using the score of that statement.

Text-independent: in this method, the speaker is enabled to repeat what statement he/she wishes and the system conclude based on that statement. This method is more difficult compared
to the first one with lower accuracy (Nancie Gunson, Diarmid Marshall, Fergus McInnes, Mervyn Jack, 2011).

In banking applications, certain capabilities are used to create a speech-oriented telephone bank software:

- Creating separate groups for research task
- Identifying WAV and MP3 formats
- Online telephone identification passively
- Identification using file in Mono and Stereo modes
- These systems are of various applications including:
  - Authentication systems like bank telephone system
  - Legal and court use for criminals identification
  - Military applications
  - Modeling methods

Gaussian mixture model is one of the well-known methods used for modeling in speaker authentication. It is a statistic model within which the viewers are enabled to observe modes and uses a normal distribution. It means that all natural phenomena use a normal distribution and the sound is not an exception and in some cases is hidden Markov model. One of the advantages of this system is modeling intricate distribution of speakers’ space. The biggest problem is here that it is non-distinguishing. Since speaker authentication is a quite distinguish-based process researches attempt to use other methods for this task. Supporting vector machine is a method used in various applications for recognizing a pattern for modelling distinguishing problems. Hence, later methods have used this method combined to Gaussian mixture model to increase authentication efficiency of systems. These methods were stronger decreasing channel effects but the sensitivity problem for channel was still present. The latest method presented in this field is the method of modelling of speakers’ space by general changes space and using probabilistic linear discriminative analysis for removing the channel. Removing two sensitivity to channel effects and to non-speaker data are among most important challenges to name some (N. Dehak, 2009).

**Challenges faced on using this technology in e-banking**

As mentioned earlier one of the main challenges in speaker authentication is changing conditions at the registration and training time. The most prevalent changes between two time intervals is changing transmission channel as well as changing sound recording channel. Recently, most of researches performed to decrease the channel effects in the field of speaker recognition. Channel changes lead to loss of recognition system accuracy in two ways. First, difference between training channel and test speech for a speaker results in less similarity of two speeches and as a result increases error of wrongness rejection. The second is when the transmission channel is similar for two different speakers and the channel affects deeply the speech, too. In this case, the similarity of two speakers’ speech with similar channels will be higher which in turn will result in wrong authentication error (Zin Ali, 2004).
Another challenge in this field is short training and test speeches. The shorter the speech is the higher the sensitivity to the expressed content and this will result in increased error. Recently, several methods have been proposed to remove this problem however it is still one of most challenging research issues. The issue of short speeches is worse in text-dependent mode. In this mode, because the content of training and test speeches is similar the short speech makes the similarity of content on phonemecontent rather than speakers’ characteristics. Hence, the error of wrong authentication increases. The reverse issue is the case for the text-independent mode. When the registration and test speeches are short their similarity will be reduced and makes the wrong authentication to be increased. However, this is originated form higher dependency on the content in short speeches. To remove the problem of content-dependency non-speaker data is primarily discussed. It is considered as all data on data vector as well as in scoring criterion that is resulted in increased error.

**Speaker authentication system**

An authentication system is consisted of two steps: registration (training) and test. In registration step, the speakers’ speech characteristics are first extracted and using those characteristics one or more models are devised for a speaker. Generated models are saved for later uses in data bases. This is the case for test step. First those characteristics are extracted and then based on authentication method they are scored and finally are rejected or accepted comparing to a threshold value (Nancie Gunson, Diarmid Marshall, Fergus McInnes, Mervyn Jack, 2011).

**Authentication system components**

Generally, an authentication system is consisted of three components. The first is speech properties extraction. Normally, various levels of speakers’ data is used. The first level indicates characteristics of speaker’s voice path and has the highest application in this field. The second level models phonological characteristics. These characteristics are most related to the way of phonemes articulations by various speakers. Other level relates to speakers’ behavioral properties including pitch frequency and its changes, tone, energy and length of phonemes (N. Dehak, P. A. Torres-Carrascoilo, D. A. Reynolds, and R. Dehak, 2011).

The second part is modeling. In this part, a model is devised for any speaker. There are various methods have been proposed for modelling classified into two categories: generative and discriminative. The final part of the system is scoring and decision making. Based on the type of modeling, there are various methods for obtaining a score. It is possible the obtained score to be normalized through various methods. Finally, a threshold value is required for decision making to compare an obtained score with. The value of the threshold may be identical for all speakers or obtained separately for every speaker. After being compared to the threshold, in case that exceeds the threshold the speaker is approved and rejected otherwise (L. D'Haro, R. Cordoba, M. Caraballo, and J. M. Pardo, 2013).
Conclusions

Reviewing literature in speaker recognition field, we concluded that using cellphone for banking affairs round the clock all week long in the field of e-banking is possible and accessibility hours are considerably longer than branch hours. It should also be noted that this application is accessible for all classes in the society without knowledge, financial and geographical limitations. Therefore, this necessity would be highly functional in e-banking. Based on researched I have performed, increasing customer satisfaction is one of causes leading to creation of this technology in future. This method is better than previous methods for elderly and physically disabled people. By this method, banks and related organizations will save time needed for customer response and financial costs. Those services enable customers to perform vast range of banking tasks through cellphone from quite simple questions to opening an account. Most of those services are automated to some extent meaning that the caller hears a series of automated messages and responses pushing keyboard buttons or his/her voice based on designed services.

Data classification for making relationship to customers and banking system in various sectors is a very essential issue which leads to time saving. Regarding the fact that guiding users through interactive voice response systems based on tone-troche time consuming, speech systems lead to less complicated relationship and time of dialogue and busy lines. People are enabled in these systems to contact any departments and to be recognized via their speech what services they need just like the case in which the customer interacts with an operator. It thus is predicted that these systems will be replaced to old systems. This application can be used in mobiles, too. Here will be problems because of speech styles including colloquial speech or different pitches or various accents available around the nation. In these cases, factors available in voice pitch frequency may be used. Moreover, there are more technical issues like quality of the entering voice from telephone lines or cellphones and frequency band with for telephone signal.

Another point is authentication using customers’ voice. This method may be used for identifying callers and increasing security of banking systems. Regarding the fact that the voice is innate with people all the time and there is no problem like being lost or stolen as well as the customer isn’t required to attend physically for identification, this method is more secure. Currently, some of Colombian, American and British banks and an American passenger trains Co. perform 56-72% of their calls through speech systems. Among those methods presented for speaker modeling the best is certainly identity vector in general changes space. Experiments indicate that the identity vector contains various data including phonemic content, lingual data, speaker data, feeling and etc. hence; in this vector there is much non-speaker data that must be investigated. The probabilistic linear discriminative method is the best method ever presented for removing the effects of channel. In this method factors related to the speaker and channel to be separated to remove the effects of channel from the final decision-making.[H. Beigi, 2011]
References


