International Journal of Circuit, Computing and Networking

E-ISSN: 2707-5931
P-ISSN: 2707-5923
Impact Factor (RJIF): 5.64
Journal's Website
IJCCN 2025; 6(2): 107-110
Received: 10-07-2025
Accepted: 15-08-2025

Maroa Essam Baker
Technical Engineering College, Northern Technical University, Kirkuk, Iraq

Corresponding Author:
Maroa Essam Baker
Technical Engineering College, Northern Technical University, Kirkuk, Iraq

Using a hidden Markov model to improve and alter automated speaker verification

Maroa Essam Baker

DOI: https://doi.org/10.33545/27075923.2025.v6.i2b.116

Abstract
The research team behind this project set out to answer several questions about autonomous surface vehicle (ASV) system design and execution. Given the advantages they provide over other biometric techniques, the development and enhancement of ASV applications is crucial. support vector machines Hidden Markov models, artificial neural networks generalized method of moments (GMMs), and combination models are the backbone of modern speaker identification systems. The efficiency of prompted text speaker verification is examined in this research using a dataset from France. In this work, a continuous speech system based on HMM has been constructed at a context-free, single mixed monophony level. The next step is to construct the client and world models using appropriate speech data. The text-dependent speaker verification method use newly-joined HMM sentences as the key text for speaker verification. During the verification stage, the normalized log-likelihood is calculated as the difference between the log-likelihoods of the client model (as enforced by the Viterbi algorithm) and the world model. The results of the verification may now be calculated thanks to a recently disclosed approach.

Keywords: Speaker, artificial neural networks, support vector machine.

Introduction
Methods that automatically recognize persons with financial operations, security, electronic banking and law enforcement, health, and counterterrorism measures are among the many applications that may be found in the modern world. The area of homeland security is characterized by a number of physiological and behavioral characteristics, some examples of which include retail sales and social services. A significant number of these firms are already using technology that is based on biometrics [1]. By using the Markov chain model, it is possible to make predictions about the probability distributions of random variables and states that are capable of symbols, tags, and words, such as those that are used to indicate the weather forecast. In order to make accurate predictions about the future, a Markov chain operates on the assumption that the most significant aspect is the present state. Prior to the current circumstance, it is impossible for anybody to make a prediction about what will occur in the future. It is necessary for an authentication system that makes use of biometrics to compare a biometric sample that has been previously registered with a biometric sample that has been recently collected [4, 5]. The collection, examination, and storage of a biometric sample for eventual comparison during the registration process are all performed. Using recognition in either identification or identification mode (the system chooses the best match from the whole enrolled population) are two instances of this [6, 7]. A person's claimed identity may be confirmed or validated using recognition in either manner. Within the realm of biometrics, approaches for speaker recognition have been considered "classic" ever since the 1970s [8, 9]. In order to determine the acoustic features of each individual speaker, speaker identification systems extract these characteristics from the spoken stream. The following qualities are represented in the following factors: [10] Anatomy refers to size forms the and geometrical of the voice lips cords, lips, teeth, lungs, velum, and tongue [11]. Learned behavioral patterns refer to the method in which an individual talks [12]. Learning refers to the manner in which an individual speaks. Within the realm of speaker recognition, speech signal processing include both identification and verification components. Through the use of speaker verification, an individual's identity may be confirmed to be that of the person they claim to be.

The purpose of a speaker identification system is to ascertain whether or not a certain person or group is associated with a particular speaker Individuals are the ones who make a claim of identify when it comes to speaker verification. Regardless of the phonetic nature of the sentence, the system will identify it while utilizing text dependent recognition. Contrarily, phrase recognition is required when text independent recognition is used, irrespective of whether the suggestion is visual or auditory.

In this study, we employ a previously established continuous speech recognizer to create a text-dependent, high-quality, cooperative speaker-friendly autonomous surface vehicle (ASV) system. It usually just takes a few simple steps to confirm a speaker's identity: the claimant talks into a microphone, and the system picks up on it. Depending on the situation, the system may reject the claim and request further speech, declare a lack of confidence, refuse or accept the claim, or make no conclusion at all.

In the ASV system, the proposed method is dependent on the textin

For the purpose of this investigation, the major focus is on the dependent-text-system that was constructed using the earlier continuous speech recognition (CSR) technology as a foundation. In what follows, you will find a more detailed description of the French CSR data collection. It uses a monophone- free single-level context-mixture hidden Markov model to generate the end output. The process of building a speaker-specific automated speech verification system consists of five steps:

a) In order to train a speaker-dependent automatic speech recognizer (SDASR), each and every phrase that is obtained from the client is used. It is necessary to carry out this enrollment process for each and every new customer.
b) In the first step of training a speaker independent automatic speech recognizer, all of the phrases that are available in the database are used as part of the training set. Client sentences are not included in the training pool. Preparing the global model for study.
c) The Viterbi forced alignment procedure is used to align all of the test phrases, Through the process of comparing the regions of overlap between the two cumulative distribution functions (CDF), the error probability may be decreased to its bare minimum (CDF). As surface area reduces, the likelihood of the ASV system producing inaccurate results diminishes as well. An experimental approach may be used to estimate the unknown density functions, which are present for all speakers (including the client).

Recognition of words for the continuous future

An older speech recognition system for the French language served as the foundation for the ASV system, which used a continuous speech recognizer. A dictionary specific to the English language. Over the course of this investigation, a database that was developed in an office environment has been used. Included in this collection are more than ten hours of read speech from eleven distinct speakers, eight of whom are male and five of whom are female. The texts include a total of 3100 phrases and over 2000 word phrases that are often used in a variety of fields, including education, sports, politics, and other fields. All of the speakers have two sets: one for training reasons and another for testing purposes.

Characteristics are extrapolated from there. For the purpose of recording the waveforms, a cardio id desktop microphone running at 30 Hz 7 kHz and using a conventional 64-bit PC sound card with 6 kHz/6-bit sampling. A 30 dB signal-to-noise ratio was used. The waveform was pre-emphasized with a coefficient of 0.621 before being parameterized using a cepstral technique on a 24 ms (HM) Hamming window. The end result was an overlap of 51%, after all these steps were taken. A total of 37 parameters have been derived from the recovery and addition of the second and first derivatives of 12 fundamental Mel-frequency cepstral coefficients (MFCCs).

Sound waves serve as the basis for the models. In order to provide a description of each French phoneme, a three-state HMM was used, and a single mixed Gaussian continuous distribution was utilized. During the process of determining the output probability, the covariance matrices are diagonal in order to reduce the amount of computations that are required. To begin, each and every model is constructed in a same manner. Following this, the Baum-Welch procedure, which is also referred to as embedded training, is used. We begin by determining the covariance and global speech mean (across all training samples), and then these parameters are used to initiate the whole set of HMMs, all of which are identical. Furthermore, a multitude of models may be distinguished by the use of embedded training. Each training phrase becomes a composite model when independent models based on a phonetic lexicon and phrase labels and are combined.

In the process of training context-free models, the following are the primary phases that are often involved:

a. Viterbi forced alignment is used in situations when the training lexicon has a large number of pronunciations in order to choose the candidate who has the greatest alignment score.

The process of re-estimating the Baum-Welch parameter requires between four and six iterations to be finished.

b. The initialization of each and every HMM is identical.

When it comes to composite models, re-estimating the Welch -Baum- parameter requires between four and six iterations (with a convergence criteria of 0.02 in log-likelihood).

The results and the discussion

It was decided that the subject of the investigations would be a speaker chosen at random from the pool of existing speakers. A client model for this speaker, which was designated as SD-CSR3, was developed with the assistance of a speaker-dependent continuous speech recognizer. After that, using a speaker-independent continuous speech model, a global model was developed for the other speakers included in the database.

CSR-SI stands for system of recognition. Neither a client model nor a world model differs from one another in terms of their structural and acoustical qualities; the only thing that differentiates the two models is the training data.

During the verification stage, the phrase that is and aligned using a Viterbi forced alignment technique. After that, in (1), we get the normalized score for both systems. Figure 1 shows the ASV's architecture. When the normalized score is subtracted from the threshold, a decision is made based on the result. Table 1 displays some of the results for the test phrase "S0001," which was spoken by both the client (the

real customer) and an imposter (speaker #1) who wore the identical clothes as the real customer (male). Client and imposter phrase normalized scores differ.

The client receives a score that is more than zero, whilst the impostor receives a score that is lower.

For the purpose of determining the capabilities of the ASV system, a normalized score is created for each of the test phrases. One of the factors that is considered while establishing the grade is the average word score for a particular test phrase. Figure 1 has the ability to display the normalized score that was obtained from the test phrases. Every collection of 200 speaker phrases includes phrases from each and every one of the individuals that are a part of the group. Taking into consideration the third speaker, who is the customer, the phrase indices range from 702 to 800, for instance? Given that the phrases provided by the client get a better score than those provided by other speakers, we are able to utilize this as a factor for deciding whether or not to accept a proposal.

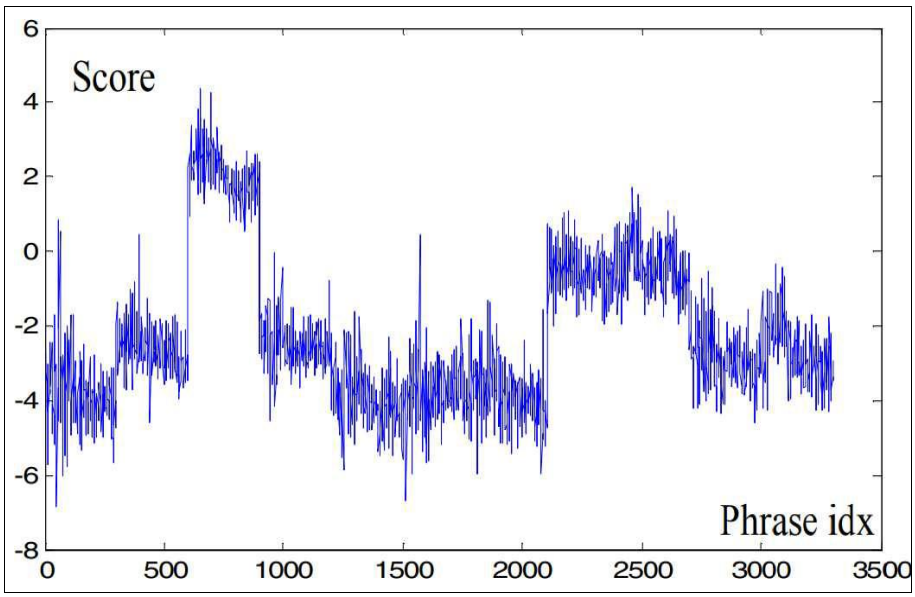

Fig 1: A calculation of the average score for each of the test phrases

Figure 2 shows the proper way to determine the FAR (FAR) and FRR (FRR) values. Limiting the range of variation that the threshold T can allow are the lowest and highest scores that were achieved over the complete phrase test set.

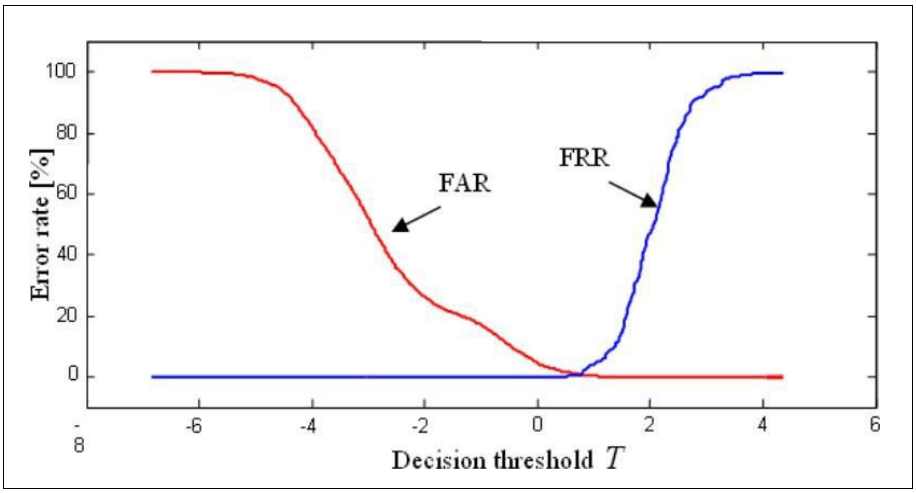

Fig 2: Both the FRR and the FAR

Presented with the criterion are the error rates and thresholds that are associated with each of the criteria. It is feasible to see that a higher threshold (R = 2.65), which is necessary in order to obtain a false acceptance rate of zero, leads to a significant proportion of false rejections (FRR percent = 1). In light of the fact that the system ought to prevent phonies from passing through, it is probable that this is the case with access apps. It is possible that the number of attempts will be increased in order to lower the FRR. For the purpose of determining how well each speaker model performed, it was put through a series of tests in which it was compared to the phrases used by the other speakers. According to the decision, the terms of the client will not be included into the global model.

Table 1: Threshold results

Criterion	Threshold	FRR [%]	FAR [%]
Minimum FRR	0.51	1.45	0.0
Minimum FRR × FAR	0.76	0.5	0.80
Minimum FAR	1.75	0.0	42

Conclusion

The purpose of this research is to present an automatic speech verification system that is based on continuous speech speakers.

The training step is responsible for the acquisition of two recognizes: one recognize is dependent on the speaker for the client model, while the other recognize is independent of the speaker for the world model. Every single one of the two recognizes makes use of the identical collection of parameters and has the similar composition. During the phase of verifying the input phrase, a Viterbi algorithm with forced alignment is used. A comparison of the normalized acoustic score to an acceptance threshold is going to be necessary in order to determine whether or not it satisfies the standards.

The experimental results of this technology demonstrate that it is capable of achieving error rates that are lower than one percent. When the search space constraint is taken into mind, the use of a forced alignment strategy minimizes the amount of time required for recognition and, as a result, the amount of money spent on computing. In the studies, a central processing unit (CPU) operating at 1.5 GHz was used, and the average processing time was less than one second. Despite the fact that the verification device is very affordable to install, a large amount of resources are needed during the enrollment step in order to construct the client's speaker-dependent recognizer from the ground up. It is possible to solve this issue by using conventional adaptation techniques, such as MLLR or MAP, in order to adjust the parameters of the world model appropriately for the new speaker.

References

1. Flanagan JL. Speech analysis, synthesis and perception. Vol. 3. New York: Springer Science & Business Media; 2013.
2. Cummins N, Baird A, Schuller BW. Speech analysis for health: current state-of-the-art and the increasing impact of deep learning. Methods. 2018;151:41-54.
3. Cummins N, Scherer S, Krajewski J, Schnieder S, Epps J, Quatieri TF. A review of depression and suicide risk assessment using speech analysis. Speech Communication. 2015;71:10-49.
4. McAulay R, Quatieri T. Speech analysis/synthesis based on a sinusoidal representation. IEEE Transactions on Acoustics, Speech, and Signal Processing. 2003;34(4):744-754.
5. Astrahan MM. Speech analysis by clustering, or the hyperphoneme method. Stanford (CA): Stanford Artificial Intelligence Laboratory; 1970. Report No.: AIM-124.
6. Krishnamurthy A, Childers D. Two-channel speech analysis. IEEE Transactions on Acoustics, Speech, and Signal Processing. 2003;34(4):730-743.
7. Busso C, Lee S, Narayanan SS. Using neutral speech models for emotional speech analysis. In: Proceedings of Interspeech; 2007 Aug. p. 2225-2228.
8. Maragos P, Kaiser JF, Quatieri TF. Energy separation in signal modulations with application to speech analysis. IEEE Transactions on Signal Processing. 2002;41(10):3024-3051.
9. Fulop SA. Speech spectrum analysis. New York: Springer Science & Business Media; 2011.
10. Ousidhoum N, Lin Z, Zhang H, Song Y, Yeung DY. Multilingual and multi-aspect hate speech analysis. arXiv preprint arXiv:1908.11049. 2019.
11. Hughes GW, Hemdal JF. Speech analysis. Lexington (MA): MIT Lincoln Laboratory; 1965. Report No.: TREE-659.