International Journal of Analytical, Experimental and Finite Element Analysis
Volume 10, Issue 4, December 2023, pp 119 - 125
https://doi.org/10.26706/ijaefea.4.10.20231904

Analysis of Speaker Verification Method Based on Hidden Markov Models for Continuous Speech System

Mahdi Fadil Khaleel

Community Health Techniqes Department, Kirkuk Technical Institute, Northern Technical University, Iraq.
Email: mahdi.fadil@ntu.edu.iq

Abstract: This paper aim to the effectiveness of speaker verification using prompted text. The progress and enhancement of ASV applications have significant implications, particularly considering their advantages in comparison to alternative biometric methodologies., support vector machines (SVM), Hidden Markov models (HMM), the generalized method of moments (GMM), artificial neural networks (ANN), and combination models are only some of the statistical models used by modern speaker recognition systems. Using a dataset collected in Turkish. The goal of this work was to create a continuous speech system using Hidden Markov Models (HMM) on a single mixed monophonic level, independent of any surrounding environment. Subsequently, appropriate speech data is used in the construction of both the client and world models. The text-dependent speaker verification method employs sentence Hidden Markov Models (HMMs) that have been concatenated for the designated text in order to authenticate speakers. The normalized log-likelihood is calculated in the verification stage by comparing the log-likelihood of the client model, which is derived using the Viterbi method and the world model. It is by subtracting these two log-likelihood values that we arrive at the normalized log-likelihood. Finally, a method for evaluating verification results is shown.

Keywords: Text dependent, Turkish data set, Viterbi algorithm, Generalized method of moments.

Review – Peer Reviewed
Received: 26 October 2023
Accepted: 15 December 2023
Published: 30 December 2023

Cite this article: Mahdi Fadil Khaleel, "Analysis of Speaker Verification Method Based on Hidden Markov Models for Continuous Speech System", *International Journal of Analytical, Experimental and Finite Element Analysis*, RAME Publishers, vol. 10, issue 4, pp. 119-125, 2023.
https://doi.org/10.26706/ijaefea.4.10.20231904

1. Introduction

In contemporary society, there exists a wide array of applications for techniques aimed at the automated identification of persons, particularly in the domains of security, secure electronic banking, financial transactions, law enforcement, healthcare, and counterterrorism efforts. Many other physiological or behavioral characteristics are potentially observable in the field of homeland security, including those related to social services and retail sales. The bulk of these firms are already using technology that is based on biometrics [1][2]. The Markov chain model may be used to forecast the probability distributions of random variables, specifically states that have the potential to assume values from a given set. These collections may include many linguistic elements, including words, tags, and symbols, which are used to represent meteorological conditions. In the context of forecasting future events, a Markov chain posits that the present state has paramount significance. Antecedent to the present circumstances, the future was indeterminable. In order to authenticate a user's identity, a biometric-based authentication system necessitates the comparison of a biometric sample that has been previously recorded with a freshly obtained biometric sample [3][4]. During the registration process, a biometric sample is obtained, subjected to analysis, and then stored for future reference and comparison. When set to identification, the system will find the best possible match among all of the people who have signed up. As a "classic" biometrics technique dating back to the 1970s, speaker recognition has been around for quite some time.

Speaker identification systems extract the acoustic features of each speaker from the spoken stream. The aforementioned attributes are exemplified in: i) The study of anatomy encompasses the examination of the geometric and dimensional characteristics of several vocal organs, including the tongue, lungs, teeth, lips, vocal cords, and velum. ii) The acquisition of acquired behavioral patterns plays a significant role in shaping an individual's speaking style. Additionally, the process of learning itself influences the way in which one talks[5][6][7].

Speaker recognition encompasses both verification and identification processes within the domain of speech signal processing. The term "speaker verification" refers to a process that verifies a person's stated identification in order to ensure they are who they say they are. One purpose of speaker identification technology is to verify the individual or group affiliation of a speaker. In the context of speaker verification, an individual asserts a claim of identity. When applied to text-based identification, the system may identify a sentence without regard to its phonetic components. On the other hand, with text-independent recognition, it doesn't matter if the phrase is offered visually or aurally; the system must be able to recognize it [8][9].

This paper uses a preexisting continuous speech recognition system to develop a system for autonomous surface vehicles (ASVs) that relies on text dependency and cooperative speakers. Common methods for verifying the identification of a speaker include a few simple steps, such as having the person claiming their identity say a predetermined sentence into a microphone. Subsequently, the system evaluates the utterance and determines whether to accept or reject it, or alternatively, expresses a lack of confidence and requests additional speech data before rendering a final decision [10][11].

2. Text-Dependent Proposed Method for The ASV System

The major focus of this study is the Examining the text-dependent system built on top of current continuous speech recognition (CSR) technology. Following this introduction, a detailed examination of the breadth of the Turkish CSR dataset will be presented. An unsupervised hidden Markov model (HMM) with a single mixture, performing at the monophony level, generates the final output. There are a total of five steps that must be taken in order to develop a personalized automated speech verification system for a certain user (the client).

i. To begin, while a speaker independent automatic speech recognizer (SIASR) is being trained, only phrases from the database are utilized as training data, and client-provided sentences are not taken into account. The process of training the global model is being conducted.

ii. All of the client-collected phrases are used in the creation of a speaker-dependent automatic speech recognizer (SDASR). The process of registering new clients involves the execution of this activity.

iii. The alignment of test phrases is achieved by the Viterbi forced alignment procedure, resulting in the generation of two acoustic scores for each phrase (observation) [13].

   a. The log-likelihood of the SD-ASR logP (O|λSD) has been computed.

   b. The log-likelihood of the observed data given the SI-ASR model parameters, denoted as λSI, has been computed [14].

The observation sequence of the phrase, denoted as O, yields two speaker-specific Hidden Markov Models (HMMs), namely λSI and λSD.

iv. To improve the stability of recognition, the acoustic score may be normalized by computing the normalized score, as shown.

$$L(O) = \log P\,(O|\lambda SD) - \log P(O|\lambda sr) \quad (1)$$

v. In order to gauge how well a phrase set is doing, we must first ascertain whether or not it has a high false rejection rate (FRR) and low false acceptance rate (FAR) within a certain threshold T range.

The system for automatic voice verification that was tested is shown in Figure 1. Under the null hypothesis H_0, the conditional density function of the scores of the other speakers is denoted by $p_A(z|H_0)$, whereas under the alternative hypothesis H_1, the conditional density function of the claimed speaker's scores is denoted by $p_A(z|H_1)$. When both

speakers' conditional density functions, $\lambda_A(z)$, are known, the Bayes test, which assumes that A and B have similar misclassification costs, uses the likelihood ratio.

$$\lambda_A(z) = \frac{p_A(z|H_0)}{p_A(z|H_1)} \tag{2}$$

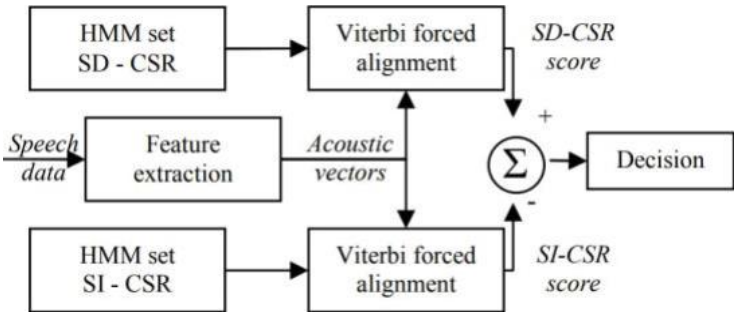

Figure 1. The ASV system design schematics

By conducting a comparison of the overlapping regions between two cumulative distribution functions (CDFs), the likelihood of error is minimized to its lowest possible value (CDF). The probability of inaccuracies in the ASV system diminishes as the surface area decreases. Experimentation may be used to create educated guesses about unknown density functions, such as those associated with the client speaker and the other speakers.

i. Using the acoustic scores produced by the client speaker's own model, the conditional probability density function $p_A(z|H_1)$ is derived for client speaker A.

ii. The scores of other speakers are computed using the speaker A model, and this serves as the probability density function (pdf) applied to impostors, denoted as $p_A(z|H_0)$.

Therefore, the likelihood ratio $\lambda_A(z)$ may now be computed. Once the threshold value T has been chosen, the classification process resumes, focusing on establishing the appropriate categorization. Additionally, the decision rule undergoes a modification.

$$\lambda_A(z) = \frac{\geq T \text{ choose } H_0}{\leq T \text{ choose } H_1} \tag{3}$$

The criteria used for the determination of threshold T is as follows:

iii. The a priori likelihood of the user being an imposter is compared against the probability of being the real speaker to calculate the value of T, which is then used to achieve a low error performance.

iv. The process of choosing the appropriate value for T to fulfill a certain FA or facial recognition (FR) criterion.

v. Modifying the temperature, denoted as T, until the desired ratio of FR (fuel consumption rate) to FA (air consumption rate) is attained.

The threshold value, represent as T, has the probability to be tailored to the specific requirements of each individual client within a certain database. Alternatively, it is also possible to dynamically modify the boundary value, taking into consideration a range of parameters. Errors in false rejection (FR) and false acceptance (FA) rates exhibit a spectrum of potential values. Put simply, when the threshold increases, the quantity of false acceptances and false rejections also increases. Mistakes related to false acceptance (FA) and false rejection (FR) are often mitigated by a process of balancing, which is determined by a predetermined threshold value.

3. Long-Term Word Recognition

The ASV system employs a continuous speech recognizer that is based on a pre-existing Turkish language speech recognition system. An English lexical resource that encompasses synonyms and associated words. This study utilized a database that was created within a professional office setting. This set consists of about 10 hours of spoken talk, featuring contributions from a total of eleven individuals, including eight males and five females. The texts consist of a

thorough collection of 4100 phrases and over 3000 regularly used terms sourced from several academic disciplines, encompassing education, sports, politics, and other pertinent topics. Each speaker has access to two sets of data: one for training and one for testing.The action of extrapolating characteristics. A desktop microphone optimized for cardiovascular identification was used to capture the waveforms; its frequency response was measured to be between 32 Hz and 8 kHz. In addition, a standard 32-bit PC sound card with an 8 kHz/8-bit sampling rate was employed throughout the recording process. As a whole, the recording procedure resulted in a signal-to-noise ratio (SNR) of 31 dB. We pre-emphasized the waveform with a coefficient of 0.91 before subjecting it to cepstral parameterization. The experiment was carried out using a 25 ms (HM) Hamming window, and the end result showed an overlap rate of 41%. All 13 major Mel-frequency cepstral coefficients (MFCCs) and their first and second derivatives have been calculated. Combining the previously stated derivatives yields 37 adjustable variables.

Models based on acoustic waves. Each Turkish phoneme was individually characterized using a three-state Hidden Markov Model (HMM) with a single mixed Gaussian continuous distribution. The covariance matrices are constructed such that they are diagonal in order to make the most efficient use of processing resources for calculating the probability of the outputs. Following a standard method of model construction, the Baum-Welch approach (also known as embedded training) is applied to all of them. Aggregating values across all training utterances yields the global speech mean and covariance. The whole collection of Hidden Markov Models (HMMs) is then initialized with these combined values, guaranteeing that each HMM starts with the same information. To distinguish one model from another, embedded training is also used.

In what follows, take a look at the main steps of how context-free models are trained.

i. All Hidden Markov Models (HMMs) have the same starting point because of how they are initialized.

ii. For composite models, reestimating the Baum-Welch parameter generally takes between 4 and 6 iterations, using 0.01 as the assumed convergence criteria in log likelihood.

iii. When there are several possible pronunciations for a word in the training lexicon, the Viterbi forced alignment method is employed to choose the one with the highest alignment score.

iv. Re-estimating the Baum-Welch parameter is a multi-step procedure that usually takes between four and six tries.

4. Discussion And Results

The study subject was a randomly selected speaker from among those who had agreed to participate. For the purpose of developing a client model for this speaker (SD-CSR3), a speaker-dependent continuous speech recognizer was used. Following this, a speaker-independent continuous speech recognition system (CSR-SI) was used to construct a global model for the remaining speakers in the database. Similar structural and acoustic properties may be seen in both the client model and the global model; the primary difference between the two lies in the data utilized for training.

A Viterbi forced alignment approach is used to align the phrase under evaluation during the verification phase. The normalized score is then calculated using equation (1) for both systems. Figure 1 depicts the ASV's overall architectural layout. Then, the difference between the normalized score and the cutoff is used to reach a conclusion. Take a look at the data in Table 1 for the phrase "S0001," which was said by both the client (hereafter referred to as the customer) and an imposter (hence referred to as speaker #1), both of whom wore clothes that were comparable to those worn by men. Normalized scores for both the client and impostor phrases show a disparity, with the client's score being above zero and the imposter's score being below it.

Each test phrase is given a normalized score that is used to evaluate the ASV system. Each phrase on the exam is evaluated using its mean word score to determine its final grade. Figure 2 displays the normalized score based on the test sentences. Each collection of 300 speaker phrases is a collaborative effort amongst all members of the speaker community. When considering the third speaker (the client), it is important to note that the phrase indices display a range of values, namely from 601 to 900. The fact that customer phrases tend to have higher mean scores than phrases from other speakers shows that this may play a role in determining whether or not a proposal is accepted.

Table 1. Both the imposter and the customer paid acoustic degree prices for the same phrases.

The logarithmic (acoustic) score Z customer Imitation (Speaker)						
Phrase	**N score**	**ST score**	**SD score**	**Norm score**	**SI score**	**SD score**
gel	6.77	−61.42	−69.20	−1.84	−66.96	−64.12
pardon	3.10	60.24	−64.35	−0.77	56.46	−55.58
Betek	2.10	−58.20	−61.31	−6.52	−66.79	−59.58
Bey	3.79	−64.01	−68.81	−2.39	−65.66	−62.2
Kan	5.15	−60.22	−66.22	−2.25	−63.87	−60.61
ARI	1.09	−63.11	−64.22	−12.65	−73.27	−59.61

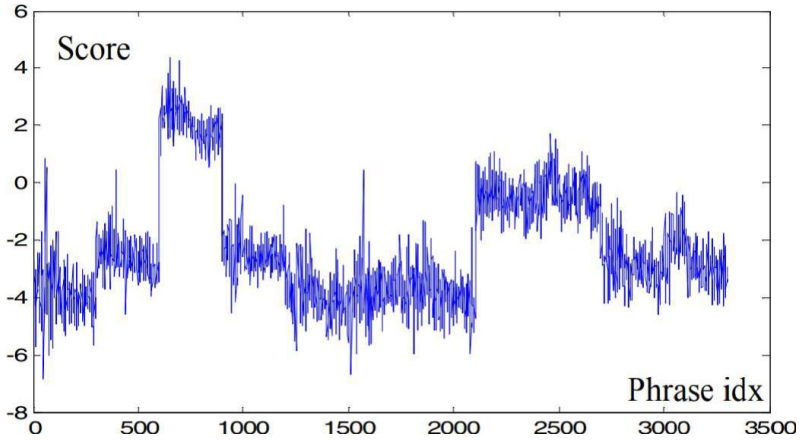

Figure 2. A weighted average was calculated for each test sentence.

The following measures are used to evaluate the performance of a system:

1. The objective of this analysis is to examine the outcomes of the Automatic Speech Recognition (ASV) system for speaker #3, specifically focusing on comparable speaker phrases. The aim is to estimate the probability function $P3(z\lambda H1)$.
2. An estimate of the probability function $P3(z\lambda H0)$ was calculated based on Speaker 3's rating of the other speakers' utterances. The probability function distributions (pdf) for speakers $P3(z\lambda H0)$ and $P3(z\lambda H1)$ may be computed using the information supplied.
3. The judgment threshold (T) is used to calculate the false rejection rate (FRR).

$$FRR = (T) = \frac{N_c(T)}{N_{TC}} \tag{4}$$

When the threshold value T is greater than or equal to the total number of client phrases, NC(T), the ASV (Automated Speech Verification) system rejects client phrases.

4. If you have a threshold for making judgments (T), you may calculate the FAR using this formula:

$$FRR = (T) = \frac{N_1(T)}{N_{Tr}} \tag{5}$$

The value of T indicates the cutoff point at which the Automatic Speaker Verification (ASV) system accepts or rejects fake utterances. Total number of impostor phrases is denoted by NTI, whereas the number of phrases accepted by the ASV system at a certain threshold, T, is denoted by N1(T).

5. Figure 3 displays the False Acceptance Rate (FAR) and the False Rejection Rate (FRR), both of which you must calculate. The lowest and highest scores attained across the whole phrase test set determine the bounds for the threshold T.

6. Figure 3 displays the False Acceptance Rate (FAR) and the False Rejection Rate (FRR), both of which you must calculate. The lowest and highest scores attained across the whole phrase test set determine the bounds for the threshold T.

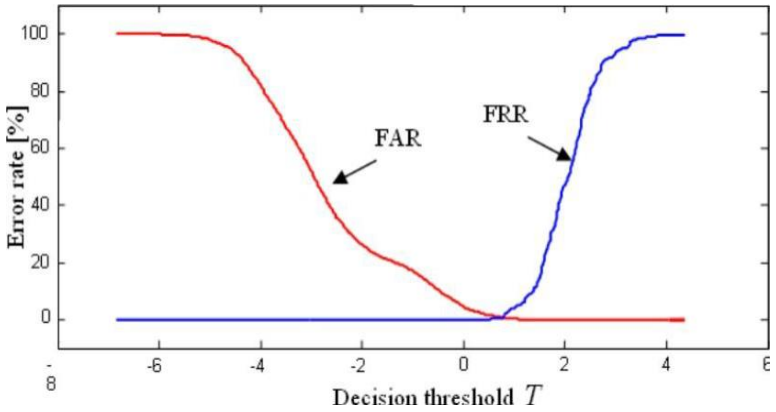

Figure 3. The FRR and FAR together

The calculation of T may be performed using either the False Acceptance Rate (FAR) or, or a combination of both. Table 2 displays a comprehensive compilation of many criteria, along by their corresponding error rates and thresholds. It is evident that in order to get a false acceptance rate of 0, a higher threshold value (T = 1.75) is necessary, which therefore leads to a significant proportion of false rejections (FRR percent = 1). It is conceivable that this may be the situation for access applications, given the system is designed to prevent fraudulent individuals from gaining entry. Increasing the number of attempts might potentially be used as a strategy to mitigate the false rejection rate (FRR). In order to evaluate the performance of each speaker model, a comparative analysis was conducted by testing them against the phrases of other speakers. The decision was made to refrain from incorporating the client's terms into the global model.

Table 2. Limiting, intermediate, and extreme measures used in deciding decisions

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

5. Conclusions

In the current study, a continuous speech recognizer is used to implement an automated speech verification system. The training phase results in the acquisition of two forms of recognition: speaker-dependent recognition for the client model and speaker-independent recognition for the world model. Both recognition systems are consistent in their structure and exhibit the same properties. In the phase of checking input phrases, a Viterbi algorithm is used to impose an alignment. By comparing the normalized acoustic score to the acceptance threshold, we may ascertain whether or not the score meets the requirements.

The empirical findings of this technological advancement illustrate its capacity to achieve error rates that are less than 1%. The implementation of a forced alignment technique within voice recognition systems leads to a decrease in the time required for recognition and thus reduces the computational expenses involved. This phenomenon can be attributed to the consideration of the constraints inside the search domain. The experiments employed a central processing unit (CPU) with a clock speed of 1.5 GHz, demonstrating an average processing duration of less than one

second. Although the verification device is considered cost-effective, a significant amount of resources is required during the enrollment phase to build the client's speaker-dependent recognizer from scratch. One possible approach to tackle this problem entails employing well-established adaptation techniques, such as Maximum A Posteriori (MAP) or Maximum Likelihood Linear Regression (MLLR), to modify the parameters of the world model in order to fit the specific traits of the new speaker.

References

[1] Zeinali, H., Sameti, H., & Burget, L. (2017). Text-dependent speaker verification based on i-vectors, neural networks and hidden Markov models. Computer Speech & Language, 46, 53-71.

[2] Zeinali, H., Sameti, H., & Burget, L. (2017). HMM-based phrase-independent i-vector extractor for text-dependent speaker verification. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 25(7), 1421-1435.

[3] Petrovska-Delacrétaz, D., & Khemiri, H. (2017, February). Unsupervised Data-driven Hidden Markov Modeling for Text-dependent Speaker Verification. In International Conference on Pattern Recognition Applications and Methods (Vol. 2, pp. 199-207). Scitepress.

[4] Olsson, J. (2002). Text dependent speaker verification with a hybrid HMM/ANN system. MASTER These Signal processing group, Uppsala University.

[5] Kadhim, I. B., Nasret, A. N., & Mahmood, Z. S. (2022). Enhancement and modification of automatic speaker verification by utilizing hidden Markov model. Indonesian Journal of Electrical Engineering and Computer Science, 27(3), 1397-1403.

[6] Hassan, M. D., Nasret, A. N., Baker, M. R., & Mahmood, Z. S. (2021). Enhancement automatic speech recognition by deep neural networks. Periodicals of Engineering and Natural Sciences, 9(4), 921-927.

[7] Nasret, A. N., Noori, A. B., Mohammed, A. A., & Mahmood, Z. S. (2021). Design of automatic speech recognition in noisy environments enhancement and modification. Periodicals of Engineering and Natural Sciences, 10(1), 71-77.

[8] Gemello, R., Mana, F., & Mori, R. D. (2005). Non-linear estimation of voice activity to improve automatic recognition of noisy speech. In Ninth European Conference on Speech Communication and Technology.

[9] Shahina, A., Yegnanarayana, B., & Kesheorey, M. R. (2004, October). Throat microphone signal for speaker recognition. In Proceedings of ICSLP.

[10] Aibinu, A. M., Salami, M. J. E., & Shafie, A. A. (2012). Artificial neural network based autoregressive modeling technique with application in voice activity detection. Engineering Applications of Artificial Intelligence, 25(6), 1265-1276.

[11] Kepuska, V. Z. (1991). Neural networks for speech recognition applications.

[12] Amrouche, A., Debyeche, M., Taleb-Ahmed, A., Rouvaen, J. M., & Yagoub, M. C. (2010). An efficient speech recognition system in adverse conditions using the nonparametric regression. Engineering Applications of Artificial Intelligence, 23(1), 85-94.

[13] Renals, S., McKelvie, D., & McInnes, F. (1991, April). A comparative study of continuous speech recognition using neural networks and hidden Markov models. In Acoustics, Speech, and Signal Processing, IEEE International Conference on (pp. 369-372). IEEE Computer Society.

[14] Karray, L., & Martin, A. (2003). Towards improving speech detection robustness for speech recognition in adverse conditions. Speech Communication, 40(3), 261-276.