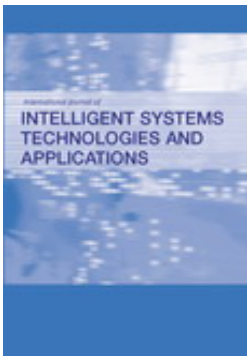
INDERSCIENCE PUBLISHERS
Linking academia, business and industry through research



International Journal of Intelligent Systems Technologies and Applications

ISSN online: 1740-8873 - ISSN print: 1740-8865
https://www.inderscience.com/ijista

Boosting speech recognition performance: a robust and accurate ensemble method based on HMMs

Samira Hazmoune, Fateh Bougamouza, Smaine Mazouzi, Mohamed Benmohammed

DOI: 10.1504/IJISTA.2024.10060581

Article History:

Received:	13 July 2023
Last revised:	07 September 2023
Accepted:	19 September 2023
Published online:	05 February 2024

Boosting speech recognition performance: a robust and accurate ensemble method based on HMMs

Samira Hazmoune*, Fateh Bougamouza and Smaine Mazouzi

Department of Computer Science,
Faculty of Sciences,
University of 20 Août 1955-Skikda,
Skikda, Algeria
Email: s.hazmoune@univ-skikda.dz
Email: b.bougamouza@univ-skikda.dz
Email: s.mazouzi@univ-skikda.dz
*Corresponding author

Mohamed Benmohammed

Department of Software Technologies and Information Systems,
Faculty of New Technologies of Information and Communication,
University Constantine 2,
Constantine, Algeria
Email: mohamed.benmohammed@univ-constantine2.dz

Abstract: In this paper, we propose an ensemble method based on hidden Markov models (HMMs) for speech recognition. Our objective is to reduce the impact of the initial setting of training parameters on the final model while improving accuracy and robustness, particularly in speaker independent systems. The main idea is to exploit the sensitivity of HMMs to the initial setting of training parameters, thus creating diversity among the ensemble members. Additionally, we perform an experimental study to investigate the potential relationship between initial training parameters and ten diversity measures from literature. The proposed method is assessed on a standard dataset from the UCI machine-learning repository. Results demonstrate its effectiveness in terms of accuracy and robustness to intra-class variability, surpassing basic classifiers (HMM, KNN, NN, SVM) and some previous works in the literature including those using deep learning algorithms such as convolutional neural networks (CNNs) and long short-term memory (LSTM).

Keywords: speech recognition; inter-speaker variability; robustness; accuracy; HMM; multiple modelling; ensemble methods; diversity.

Reference to this paper should be made as follows: Hazmoune, S., Bougamouza, F., Mazouzi, S. and Benmohammed, M. (2024) 'Boosting speech recognition performance: a robust and accurate ensemble method based on HMMs', *Int. J. Intelligent Systems Technologies and Applications*, Vol. 22, No. 1, pp.41–76.

Biographical notes: Samira Hazmoune received a Magister's degree in Computer Sciences from the University 20 Août 1955-Skikda in 2009 before pursuing a Doctoral degree at the same university, which she obtained in 2021.

She is a teacher-researcher since 2010. Her research focuses on artificial intelligence, with a particular interest in machine learning, pattern recognition and natural language processing. She has published several scientific articles in international journals and conferences, highlighting the significance of her research work in academia.

Fateh Bougamouza received a Magister's in Computer Sciences from the University of Constantine 2 – Abdelhamid Mehri in 2009 before pursuing a Doctoral degree at the same university, which he obtained in 2021. He is a Teacher-researcher since 2011. His research focuses on artificial intelligence, with a particular interest in deep learning, pattern recognition and natural language processing. He is currently an Associate Professor with the Computer Science Department, University of 20 Août 1955 – Skikda, Algeria.

Smaine Mazouzi received his MS and PhD in Computer Science from the University of Constantine, in 1996 and 2008, respectively. He is currently an Associate Professor with the University of 20 Août 1955-Skikda. His fields of interest are artificial intelligence, multi-agent systems, computer vision, cybersecurity, and machine learning.

Mohamed Benmohammed received his BSc degree from the High School of Computer Science (CERI) Algiers, Algeria, in 1983 and PhD in Computer Science from the University of Sidi Belabbes, Algeria, in 1997. He is currently an Associate Professor with the Department of Software Technologies and Information Systems, Faculty of New Technologies of Information and Communication University Constantine2, Constantine, Algeria, where he is also the Head of team in the LIRE Laboratory. His research interests are microprocessor architecture, embedded systems and real-time applications.

1 Introduction

Since the end of the fifties, several researches in automatic speech recognition (ASR) have been carried out by exploring different aspects of acoustic analysis and classification, with the main goal of improving the accuracy and robustness in order to get as close as possible to an ideal system. Various innovations and techniques have emerged over the decades to achieve this goal, including the implementation of deep learning architectures like deep neural networks (DNN) (e.g., Nassif et al., 2019; Cui et al., 2020), recurrent neural networks (RNN) (e.g., Oruh et al., 2022; Islam et al., 2019), and convolutional neural networks (CNNs) (e.g., Palaz et al., 2019; Pardede et al., 2023) for robust feature extraction. Other strategies include the integration of large-scale language models based on transformers (Vaswani et al., 2017) such as BERT (Devlin et al., 2018) or GPT (Brown, et al. 2020; Radford et al., 2018, 2019) to enhance contextual understanding, use of noise reduction and enhancement methods, transfer learning (e.g., Kunze et al., 2017; Qin et al., 2018; Wang and Zheng, 2015), multimodal fusion (eg., He et al., 2023; Song et al., 2023; Pan et al., 2022), speaker adaptation (e.g., Geng et al., 2022; Deng et al., 2023), and conduction of rigorous robustness evaluations under varied conditions. These collective strategies fuel the continuous evolution of ASR research, bringing us closer to the realisation of highly reliable and precise speech recognition systems. Unfortunately, despite the great progress made in the field, we are still far from having a natural interaction between man and machine, and many problems

are only partially resolved. These problems, which the main applications of ASR suffer from, can be classified into two main categories: Those related to the characteristics of the speech signal and to the conditions for taking measurements, and those related to the analysis and classification techniques. Among the problems of the first category, which considerably affect the quality of recognition systems in terms of accuracy and robustness, is the intra-speaker and inter-speaker variability. This variability can result from different sources, mainly related to the state of the speaker (physiological, psychological, social or even cultural) and to his environment (noise, disturbances, material and measurement conditions). The degree of variability varies, mainly, depending on the system type. In fact, a single-speaker system is less sensitive to data variability, compared to a multi-speaker system, which is, in turn, less sensitive than a speaker independent system, where anyone can use it. Beyond the variability of the speech signal, another problem related, this time, to classification techniques is the sensitivity of the classifiers to the initial setting of the training parameters. This is especially due to the inconsistency between training data and test data and to intra-class variability. The majority of classification methods proposed in the literature, such as artificial neural networks, support vector machines (SVM) and hidden Markov models (HMMs) are very affected by this problem.

HMMs are one of the most effective and popular methods in speech recognition field. They have become widely accepted as a standard speech recognition technique in the ASR community (Mustafa et al., 2019). The reasons why this method has become so popular are the availability of efficient training algorithms for estimating model parameters from finite sets of speech data (Rabiner and Juang, 1992, Hazmoune et al., 2018), their strong mathematical basis and ability to model series of variable length, such as speech signals. The training of HMM is generally based on the maximisation of objective functions (likelihood, conditional likelihood...) using the gradient algorithm (Levinson et al., 1983) or that of expectation-maximisation (EM) (Dempster et al., 1977). The latter is an iterative procedure that makes it possible to re-estimate the model parameters according to their current estimations so as to improve the likelihood of training data after each iteration. The problem is that this algorithm is too sensitive to the values of the initial parameters, which must be set carefully in order to ensure accurate results. In addition, it does not guarantee a globally optimal solution, because the objective functions are of non-convex nature, and therefore, subject to the problem of local maxima. The exhaustive search is otherwise impossible, because the number of local maxima is unknown. The setting of the initial parameters is generally optimised experimentally, and it is highly dependent on the training and test data, which affects the robustness and stability of the classifier. The question that arises is therefore: What initial setting or what final model should we choose in order to get as close as possible to the global maximum? To address this problem, some works in the literature have focused on optimising the initial setting experimentally or using combinatorial optimisation techniques such as genetic algorithms (e.g., Sosiawan et al., 2021; Xiao et al., 2007). Others have focused instead on optimising the final model based, generally, on model selection criteria (e.g., Biem, 2003). In both cases, the trained HMM model is unique and will not be able to efficiently model all instances of the data class given the great intra-class variability, especially in speaker-independent systems.

The aim of this work is to propose a novel ensemble method making it possible to design ASR systems that are both accurate and robust to data variability, based on HMMs as basic classifiers, and trying to alleviate the problem of intra-class variability and that

of sensitivity of HMMs to the initial setting of the training parameters. This paper also aims to empirically study the impact of each training parameter on ensemble diversity and, consequently, on its accuracy. The ultimate goal of this study is to decide which of those parameters has the most important impact on ensemble diversity and which of diversity measures is more adequate for ensemble selection. For this purpose, four pairwise and six non-pairwise diversity measures are used.

Contrary to heterogeneous ensemble methods, where some studies have been previously conducted to examine the relationship between diversity and ensemble accuracy, such as (Gilpin and Dunlavy, 2009), this work is the first complete experimental study of the relationship between the different diversity measures and the accuracy in homogeneous HMM-based ensemble methods.

Our proposed ensemble method for speech recognition has multiple potential applications in real-world scenarios. It can improve the accuracy of speech-to-text transcription systems, making them more helpful in generating closed captions for live video broadcasts or transcribing audio recordings, and also enhance interactive voice response (IVR) systems and increase customer satisfaction. The method can elevate speech recognition in automotive systems, including hands-free phone systems, entertainment, and climate control. Additionally, virtual assistants, which rely on understanding natural language commands, can become more efficient and reliable. The proposed method can also be integrated into medical transcription systems to improve the accuracy and reliability of transcriptions and medical record updates.

The rest of this paper is organised as follows: First, we present in Section 2, works related to the robustness against data variability, the sensitivity of HMM classifiers to the initial setting of training parameters, and the ensemble methods. Next, we describe in Section 3 the proposed ensemble method. In Section 4, we test the proposed approach, present and discuss results from several conducted experiments. Finally, we conclude and present direction for future work.

2 Related work

In this section, we discuss related works concerning the robustness to data variability, the sensitivity of HMM classifiers to initial setting of training parameters, and ensemble methods.

2.1 Robustness to data variability in speech recognition

Robustness is one of the most important aspects that one must take into account during the evaluation of the quality of speech recognition systems. The notion of robustness was introduced by Box (1979). It is defined as the ability of the system to remain stable in the face of data variability and environmental disturbances. In a consumer system (intended for use in a car, on a street, etc.), for example, noise robustness is the most important challenge, as the environment is greatly affected by several sources of noise. Additionally, to this challenge, a speaker independent system should not be overly sensitive to speaker variability, and to the inconsistency between training and testing data.

To improve the robustness of ASR systems to noise and variability, several solutions have been proposed in the literature. The majority of these solutions focus, essentially, on

the analysis phase either by selecting robust features in order to improve the signal to noise ratio SNR, by using noise reduction algorithms such as spectral subtraction or Wiener filtering, to pre-process the speech signal and reduce background noise before it enters the voice recognition system, by applying speaker normalisation to overcome speaker variability (Giuliani et al., 2006), or by combining several analysis methods (Khelifa et al., 2017). Other solutions have instead focused on the classification phase by combining more than one classification method, such as the work proposed in (Hazmoune et al., 2018), where a multiple modelling approach is carried out through a hybrid framework integrating HMM models in a k-NN architecture in order to boost robustness to intra-class variability. Another solution is the multi-modality approach, it consists in integrating other sources of information than those of the acoustic signal, as in audio-visual speech recognition (Noda et al., 2015; Feng et al., 2017), which combines acoustic and visual information, such as facial expressions and lip movement. This audio-visual combination can be carried out at the representation level, but also at the decision level. Another different way to boost the robustness of ASR systems, as cited in (Spalanzani, 1999), is to increase the size of the training dataset in order to have systems that can perform in multiple test situations. However, despite the increasingly immense size of the training datasets, they are still a very small sample of the set of possible variabilities of the speech signal. Therefore, even models that are trained on large datasets cannot effectively fit all possible types of variability. In Zelinka et al. (2012), a multi-model architecture was proposed to overcome the variability of the speaker's vocal effort, five speech modes were considered: Whispered, soft, normal, loud, and shouted. The training consists, for each data class, in learning a different model for each speech mode. The recognition consists first in identifying the vocal effort, decoding the speech signal by selecting the class models that corresponds to the identified vocal effort, then recognising lexical information using Viterbi decoding (Viterbi, 1967; Forney, 1973). The results obtained on an isolated words dataset showed (according to the authors) a 50% relative reduction of word error rate compared to the baseline system.

It should be noted that, despite its importance, the robustness aspect is not always considered during the evaluation of ASR systems, and even the works dealing with this problem, such as the ones that we have just cited, do not generally use metrics to measure the robustness of the proposed systems. They only use accuracy measures (error rate or recognition rate) to evaluate predictive performance. In this work, we propose using statistical measures to quantify the robustness of speaker independent ASR.

2.2 Sensitivity of HMM classifiers to initial setting of training parameters

Beside data variability, the initial training setting is another important factor that can affect the classifier sensitivity and, consequently, the system robustness. Before presenting related works dealing with this problem, we will first present the HMM definition as well as the practical application of HMMs in the context of deciphering speech signals. Furthermore, we will explore the HMM's training process according to the EM algorithm.

2.2.1 HMM definition

An HMM is a doubly stochastic process (Rabiner and Juang, 1986). It has two main properties: First, it assumes that a sequence of observations $O = (o_1, o_2, \ldots, o_T)$ is

produced by a sequence of hidden states $Q = (q_1, q_2, \ldots, q_T)$. In other words, the sequence of states that generated the sequence of observations is hidden from the observer, hence its name. Second, it is based on the Markov property which assumes that the state of the process at time t depends only on its state at time t–1. Based on this property, we can deduce that the probability $P(q_1, q_2, \ldots, q_T)$ that the process passes through the sequence of states $Q = (q_1, q_2, \ldots, q_T)$ can be calculated as follows:

$$P(q_1, q_2, \ldots, q_T) = P(q_1) * P(q_2 / q_1) \ * \ \ldots \ * P(q_T / q_{T-1}) \tag{1}$$

Formally, an HMM with N states and M discrete observation symbols which can be emitted by the states over time is defined by the triplet $\lambda = (\Pi, A, B)$, such as:

- Π is a vector with N elements $\pi_i = P(s0 = s_i)$, $1 \le i \le N$, representing the probabilities that the process starts from a given state.
- A is the transition matrix of size (N by N) containing the probabilities $a_{ij} = P(s_j/s_i)$ of passing from one state to another.
- B is the observation matrix of size (N by M). A coefficient $b_i(O_j)$ of B represents the probability that the symbol O_j is emitted by the state s_i.
- The following stochastic constraints must be verified

$$\sum_{i=1}^{N} \pi_i = 1, \tag{2}$$

$$\sum_{j=1}^{N} a_i j = 1, \qquad 1 \le i \le N, \tag{3}$$

$$\sum_{j=1}^{M} b_i(O_j) = 1, \qquad 1 \le i \le N \tag{4}$$

Note that this definition relates to discrete density HMMs. For continuous density HMMs, where there are no discrete observation symbols, the observation matrix B is replaced by the parameters of the probability law used to assess the observation probabilities. In the case of multi-Gaussian law (GMM for Gaussian mixture model), generally used in pattern recognition problems, the matrix B is replaced by the mean vectors and the covariance matrices of the Gaussian densities. Each density of probability associated with a state i is calculated by applying the formula:

$$N\left(\mu_i, \sum_i, O\right) = \frac{1}{(2\pi)^{P/2} \left|\sum_i\right|^{1/2}} e^{-\frac{1}{2}(O-\mu_i)^T \sum_i^{-1} (O-\mu_i)} \tag{5}$$

where, P is the dimension of the vector O, μ_i the mean vector of the density function associated with state i, $\left|\sum_i\right|$ the determinant of the covariance matrix of the density function associated with state i, $\sum_i^{-1}$ is the inverse of the covariance matrix of state i, and T the average number of observations per sequence.

The observation probabilities $b_i(O_t)$ are calculated as a weighted sum of the Gaussian density functions $N\left(\mu_i, \sum_i, O_t\right)$ associated with state *i*.

2.2.2 HMMs for speech recognition

HMMs serve as a foundational technology in speech recognition, characterising speech as a sequence of hidden states and using probability distributions to model the relationship between acoustic features and these states. Their technical intricacies encompass training, decoding, language modelling, and various techniques for robustness and adaptation, making them versatile tools in the field of speech recognition. The process of speech recognition using HMMs can be effectively dissected into two core phases:

- Training phase: The first step in speech recognition using HMMs is the collection of a large dataset of speech recordings along with corresponding transcriptions. Next, the speech recordings need to be pre-processed to remove any background noise and distortions such as microphone hiss, pops, clicks, and other sources of interference. Once the speech is cleaned, feature extraction techniques such as mel-frequency cepstral coefficients (MFCCs) are used to extract essential features from the speech signal.
- After the feature extraction process is complete, the transcriptions are used to segment the feature vectors into speech units, such as phonemes, syllables, or words. Using the segmented data, HMMs are trained for each speech unit using the Baum-Welch algorithm. An HMM consists of a set of states, each encoding different acoustic features of the speech signal. The Baum-Welch algorithm is used to maximise the likelihood of the HMM given the training data. The algorithm iteratively refines the probability distributions for each state, which estimates the model parameters, maximising the chance of recognising speech units in new recordings.
- Recognition (decoding) phase: The recognition phase of speech recognition using HMMs starts with the audio recording being pre-processed to remove background noise and other distortions. After the audio is cleaned, feature extraction techniques such as MFCCs are used to extract features from the speech signal. Next, a model-based deciphering approach is used to identify the most likely speech units in the audio recording.

In this phase, each speech unit is modelled using a set of HMMs. The Viterbi algorithm is then used to calculate the likelihood of each HMM given the feature sequence of the speech unit. The most likely sequence of speech units is selected, and the transcriptions of the recognised HMMs are combined to produce the recognised text. Finally, the recognised speech units or words are combined to produce the final transcription of the given speech signal. Overall, speech recognition using HMMs involves collecting a large dataset of speech recordings, pre-processing the audio data, segmenting the features, training the HMMs, and finally recognising speech units present in the new audio recording.

In the following, we explain how to train an HMM using EM algorithm.

2.2.3 HMM training by EM algorithm

Given a sequence of observations O, training an HMM, as we have seen previously, consists in re-estimating the parameters of the model λ while maximising the likelihood $P(O/\lambda)$. The maximum likelihood estimation ML is the standard method for estimating the parameters of a probabilistic model. According to this method, the model estimated from O and λ is the model λ'_{ML} such that:

$$\lambda'_{ML} = \arg\max_{\lambda} \left(\log\ P(O/\lambda)\right) \tag{6}$$

where $\log P(O/\lambda)$ is the logarithm of likelihood of the observed sequence O given the model λ. If we assume that the sequence O is generated by the sequence of hidden states Q, the maximisation of $\log P(O/\lambda)$ then amounts to maximising the following quantity:

$$\log \frac{P(O,Q/\lambda)}{P(Q/O,\lambda)} = \log P(O,Q/\lambda) - \log P(O,Q/\lambda) \tag{7}$$

So, we can write:

$$\lambda'_{ML} = \arg\max_{\lambda} (\log P(O,Q/\lambda) - \log\ P(O,Q/\lambda) \tag{8}$$

Since the sequence of states Q is hidden, it is impossible to directly solve the ML problem. Several approximate solutions have been proposed in the literature, such as the EM method (Dempster et al., 1977), the gradient method (Levinson et al., 1983), variational methods (Jordan et al., 1999) and others. We are interested here, in the EM method, which is the most commonly used. This considers the unobserved variables (the hidden states) as missing data and replaces them with their likelihood expectations.

The Baum-Welch algorithm (Baum and Petrie, 1966; Baum and Eagon, 1967; Baum, 1972), often referred to as the forward-backward algorithm, is the implementation of the EM method for HMMs. This algorithm enables a gradual re-estimation of the model's parameters over successive iterations. The process commences by initialising the HMM with initial parameter estimates, encompassing the probabilities associated with initial states (π), state transitions (A), and emissions (B). Subsequently, it undertakes a forward pass, meticulously calculating the forward probabilities (α) for each state at each time step, and simultaneously conducts a backward pass, computing the backward probabilities (β) in a recursive manner. These forward and backward probabilities are then joined iteratively to update the HMM's parameters, such as π, A, and B, in a manner that maximises the likelihood of the observed data $P(O/\lambda)$. This iterative refinement process persists until a convergence criterion is met, indicating that the model parameters have reached an optimal state. The final outcome of this procedure is a fully trained HMM, equipped with parameters that enable it to effectively capture and represent the hidden structure underlying sequential data, rendering it suitable for speech decoding, where a precise understanding of temporal dependencies and state dynamics is critical for accurate modelling and prediction.

During decoding or recognition, the goal is to find the most likely sequence of hidden states (phonemes or sub-phonemes) given the observed acoustic features. The Viterbi algorithm is often used for this purpose, efficiently finding the optimal state sequence by considering both the state transition probabilities and the observation probabilities.

2.2.4 Initialisation methods

The drawback of EM algorithm is that the final result highly depends on the initialisation step. The initial setting of HMM training has been the subject of study in several publications in various fields, such as in Yuan et al. (2019), Nathan et al. (1996), Liu et al. (2004), Moghaddam and Piccardi (2013), Clemente et al. (2012), Ge et al. (2016), Rabiner et al. (1984), Belaïd and Anigbogu (1994), Ferrer et al. (2000), Sosiawan et al. (2021), etc. The most important initial parameters that need to be carefully set in HMM training are the number of states, the initial model, the number of Gaussian densities associated with states and the number of iterations of the training algorithm. In the following, we present some initialisation methods from the literature.

2.2.4.1 States and Gaussian densities

Several ways have been proposed for setting the number of states of an HMM for each word in ASR. Levinson et al. (1983) suggested that the number of states should approximately correspond to the number of sound units of the word (phonemes). For his part, Bakis (1976) suggested choosing the number of states as being the average of the numbers of observations of the word's utterances. However, since the duration of utterances of the same word is variable, the optimal number of states (Kwong et al., 2001) can only be found by reconfiguring the model after each HMM learning.

To determine the number of Gaussians associated with states of an HMM, the author in Sankar (1998), proposed an algorithm called Gaussian Merging-Splitting (GMS). iterative Gaussian splitting and the EM algorithm are used to initialise the number of Gaussian densities in each state. Starting from a single Gaussian, Gaussian splitting is used to increase the number of Gaussians at each stage of the training until the necessary number of Gaussians is reached. Gaussian merging is performed before each Splitting operation. In the GMS algorithm, the user must specify the number of states and the maximum number of Gaussians per state. The number of Gaussians is iteratively increased using merging and splitting until the maximum number of Gaussians is reached.

The most common way to initialise parameters is empirical. It consists on running the training algorithm for all the plausible values, then evaluate the final system and choose the values that provide the best score. Other solutions based on optimisation techniques have been proposed, such as the one presented in Kwong et al. (2001), where a genetic algorithm is used to find the optimal number of states. In Bhuriyakorn et al. (2008), the authors proposed an approach for estimating the HMM topology (the number of states and transitions between states) for a phoneme recognition task. The process takes place in two stages: First, a set of appropriate topologies is constructed by combining different objective functions and topology generation methods. Second, a genetic algorithm is used as a topology selection method. This algorithm considers a global objective function and selects, for each phoneme, the best-suited topology among the candidates proposed in the previous step. Another solution is to apply the model selection criteria to choose the best final model from a large set of candidate models that differ in their structures (number of states or number of Gaussian densities). In Biem (2003), a model selection criterion called discriminative information criterion (DIC) has been proposed to optimise the HMMs topology (the number of states). This criterion is based on selecting the most discriminating models instead of models based on the 'Occam's razor' principle, as for

the Akaike information criterion (AIC) and Bayesian information criterion (BIC). We recall that this ('Occam's razor') is a principle of parsimony stipulating, in our case, that the model must be simple enough for an efficient calculation and sufficiently complex for the data to be well specified.

2.2.4.2 Initial model

The initial values of the transition matrix A and the starting probability Π, generally do not have a significant impact on the quality of the re-estimated model. It has been shown by Rabiner (1988) that the random initialisation, subject to stochastic and non-zero constraints (equations (2) and (3)), of these parameters, is sufficient to give useful reestimations in almost all cases. It is also common to use uniform initialisation of these parameters. On the other hand, for the parameters of B, experience has shown that good initial estimates are useful in the case of discrete HMMs, and are mandatory in the case of continuous density HMMs. Several initialisation methods have been proposed in the literature. The most common are the following:

- The segmental k-means method: This is the standard initialisation method proposed by Juang and Rabiner (1990), Rabiner et al. (1986), Rabiner (1988, 1989). It distributes the training data frames on the states using the k-means clustering and the Viterbi decoding algorithm.
- The flat-start method: The idea of this method is to make all the states equal (Itaya et al., 2005), all the models of the classes are then initialised with identical parameters equal to the mean overall and the variance of the training data (Clemente et al., 2012).

It is also possible to use, as an initial model, a random model or any model already available trained from appropriate data. Other less popular initialisation methods have been proposed such as the deterministic annealing EM (DAEM) method (Itaya et al., 2005; Kurata et al., 2006), the SMEM method (split and merge EM) (Han and Boves, 2006), and the multiple sequence method (Liu et al., 2004).

To our knowledge, and at present, there are no criteria for choosing between initialisation methods. However, some comparative studies have been conducted. The reported results do not allow demonstrating definitively the superiority of one method over another, and showed that there is no universal method that can be applied successfully to all application fields or on any dataset.

Despite the abundant literature on the initialisation methods of the EM algorithm and on the impact of initial setting on the final result, the problem of re-estimating a globally optimal model still remains unresolved. To deal with this problem, two categories of solutions are proposed in the literature. The first category includes solutions that seek the optimal initial parameters by applying optimisation methods such as genetic algorithms (Kwong et al., 2001). The second category, on the other hand, includes solutions that are concerned with the selection of a final model from a set of candidate models coming from different starting points. In this case, the model selection criteria are often used, we cite as an example the work (Biem, 2003). Unfortunately, since the objective functions used are optimised on validation sets where recording conditions are often different from those of the end-user environment (in particular, in the case of speaker-independent consumer systems), the trained models, in both categories, are too sensitive to data

variability, environmental disturbances, and inconsistency between training data and test data. Through the ensemble method proposed in this paper, we try to alleviate this problem and that of robustness to data variability by using a multiple Markovian modelling.

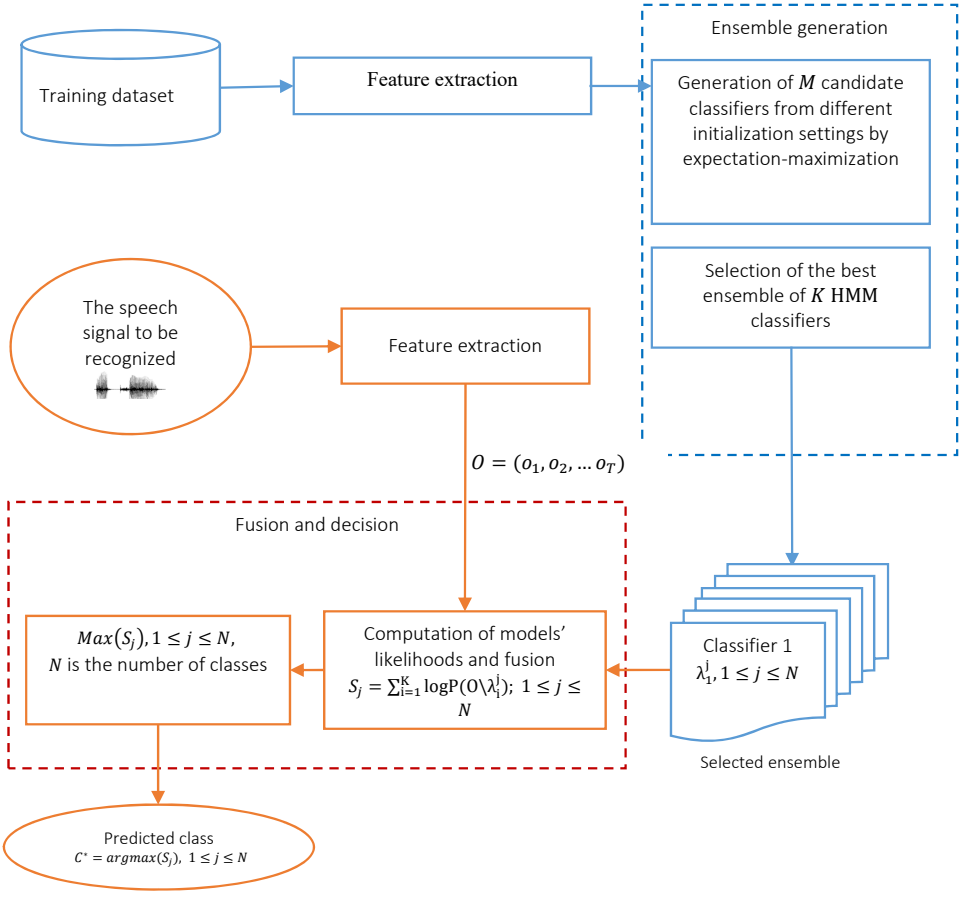
2.3 Ensemble methods

In pattern recognition area, machine learning based methods such as neural networks (NN) (Looney, 1997; Samarasinghe, 2016), SVM (Burges, 1998; Weston and Watkins, 1999; Al Dujaili et al., 2021), decision trees (Lior, 2014; Obidziński, 2021) and HMMs (Rabiner and Juang, 1986; Bougamouza et al., 2016, 2018; Hazmoune et al., 2018; Ting, 2019; Srivastava et al., 2022) have been popularly used. Although HMM classifiers may individually perform efficiently, especially in speech recognition field, more sophistication has been proposed to achieve higher performance. Often such sophistication includes use ensemble methods which aggregate heterogeneous classifiers (Kim et al., 2000; Asafuddoula et al., 2017) or homogeneous classifiers (Koerich and Poitevin, 2005; Al-Hajj et al., 2007; Guo et al., 2022) in order to approach as closely as possible to the optimal classifier performance. The effectiveness of ensemble methods can be justified by the fact that each classifier may have its own portion of data samples where it performs best (different classifiers provide different classification errors).

In the case of homogeneous classifiers (where the same classification method and algorithms are used to train all the base classifiers), ensemble methods can be classified according to the way of differentiation between ensemble members into three categories. The first one concerns methods that use different subsets of training samples to create different classifiers, such as bagging (Breiman, 1996) and boosting (Freund, 1995). In this category, training dataset is divided on several subsets, which is undesirable for generative classifiers like HMMs where considerable training data is required to get high performance. The second category includes methods using different subsets of features, such as random subspace method (RSM) (Ho, 1998). These methods are well suited for generative classifiers and applications that must deal with a limited number of training samples. The disadvantage of these methods is their high complexity, as one needs to calculate different features of the same sample for each classifier. The last category includes methods using identical training set and identical features to train all classifiers; the difference is simply done by varying parameters for which the classifier is unstable. The use of the same features for all classifiers leads to a considerable reduction in complexity and feature space compared to the second category. The approaches such as the one proposed in Hamdi and Frigui (2015) and Hazmoune et al. (2013a, 2013b) may be included in this category.

The third category is not attracting much attention in the literature compared with the first and second categories of homogeneous classifiers, but we believe that it may have a substantial impact on how much we can improve classification accuracy and robustness, mainly for generative classifiers like HMMs. As far as we know, no effort has been made to explore the impact of initial setting of HMM's training parameters on ensemble creation and diversity across the base classifiers.

Figure 1 General schema of the proposed hmm ensemble method (see online version for colours)



3 An HMM based ensemble method for robust speech recognition

The results published recently showed an impressive evolution in speech recognition field, especially in the case of isolated words with limited vocabulary. However, based on our literature review, we found that almost all works in this field are evaluated globally in terms of accuracy, ignoring the sensitivity of the system to the initial setting of training parameters. These parameters, as mentioned above, are generally optimised experimentally on the training dataset; therefore, the established models are too sensitive to data variability and inconsistency between training and test data. Moreover, in all the previous solutions of the initial setting of training parameters, only one selected value of each parameter is used, which gives a single model per class. This model may not be well suited to all samples of the class because of the great variability intra-class, mainly in speaker-independent systems.

To overcome this problem, we propose to use, instead of single modelling approach, a combination of multiple HMM with different initial settings. This allows on the one hand avoiding the problem of initial setting of HMM parameters, and on the other hand, increasing the system accuracy by combining decisions from several classifiers. Moreover, it allows increasing the system robustness by taking into account certain degrees of intra-class variability. In fact, by using multiple models per class, it is very likely that a new test sample will find, in the ensemble of generated models, a one that represents it well.

The main advantages of our ensemble method are:

1 Unlike boosting and bagging, no splitting up of the training dataset is required because all the base classifiers are trained using the whole dataset, which is strongly suited for HMMs classifier, where a large amount of data is required to achieve high performance.

2 Unlike RSM, the same features are used for the pool of classifiers, thus a considerably reduction in computational complexity.

Figure 1 illustrates a general schema of the different steps in our proposed ensemble method. Note that in this architecture, we have opted for a global modelling approach, with words as the basic units. While this approach is particularly well suited for isolated words with a limited vocabulary, it is essential to highlight that the proposed architecture is standard and adaptable to other units (such as phonemes or syllables) by substituting the word models with models of the selected unit.

3.1 Feature extraction

Feature extraction is a crucial step in speech processing, where complex audio data is transformed into a more manageable and informative format. The objective of this process is to extract relevant and discriminative features from the raw speech signal that can facilitate subsequent analysis and recognition tasks. Accurate and effective feature extraction is essential for the success of speech processing systems as it enhances the discriminative power and robustness of these systems.

Numerous distinctive features can be extracted from speech signals, each capturing specific aspects of the audio data. These encompass Mel frequency Cepstral coefficients (MFCC), which excel at capturing both spectral and temporal characteristics of speech. linear predictive coding (LPC) is another technique employed in speech processing, effectively modelling the vocal tract system as a linear filter. A modified version of LPC, known as perceptual linear prediction (PLP), incorporates principles derived from the perceptual aspects of the human auditory system, enhancing its performance. Furthermore, spectral attributes, such as spectral centroids, spectral flux, spectral roll-off, and spectral flatness, provide insights into the spectral content of the signal. Additionally, prosodic features, including pitch, duration, and energy contour, capture key elements related to intonation and rhythm in speech. These diverse feature extraction methods offer a comprehensive view of speech data, enabling a wide range of applications in speech processing and analysis. The selection of feature extraction methods depends on the specific application and dataset being used. In our work, we use the MFCC method for two main reasons. Firstly, MFCC is one of the most widely used methods in the field of automatic speech recognition (ASR), and numerous studies have shown its

effectiveness for extracting features. Secondly, due to the limitations of the dataset used in our work, we are unable to apply other features. The dataset includes only MFCC coefficients extracted from the speech signals, and audio files are not available. This will not be an issue, as our research is primarily centred around the classification phase, rather than the acoustic analysis phase. Additionally, all the systems studied in this work are based on the same MFCC features. For this reason, we will restrict our explanation to the MFCC process.

The MFCC method is based on the human auditory system's ability to discriminate sounds of different frequencies. It is effective in extracting relevant features from speech signals and converting them into a sequence of feature vectors. The process includes several steps such as pre-emphasis, framing, windowing, FFT, Mel filter bank analysis, and DCT, which work together to extract spectral and temporal characteristics of the speech signal. The pre-emphasis step compensates for the loss of higher frequency components during recording, while framing splits the speech signal into short-term segments. The windowing process applies a window function, such as the Hamming window, to reduce spectral leakage during the FFT. The Mel filter bank analysis converts the magnitude of the spectrogram to a logarithmic scale using a set of triangular filters. Finally, the DCT provides an efficient representation of the Mel filter bank output in the form of cepstral coefficients. These steps collectively transform the speech signal into a sequence of feature vectors containing coefficients that represent the spectral and temporal characteristics of the audio, making it more suitable for speech recognition tasks. In addition, Dynamic MFCC features, which are computed as the first and second derivatives of the cepstral coefficients, can be included for enhanced performance.

3.2 Generation of the ensemble

Generating ensembles of classifiers requires building individual classifiers different from each other. These do not necessarily have to be the best performing classifiers but should, when combined, provide better performance than the best performing classifier in the ensemble. In this section, we present the proposed ensemble creation method. It takes place in two steps: First, a large set of candidate classifiers, which differ in one of the parameters of the initial configuration of the training algorithm, is created using the same training dataset and the same set of features. In the second step, the best subset of candidate classifiers in terms of accuracy and diversity is selected.

3.2.1 Generation of candidate classifiers

Once the acoustic analysis is done in order to extract the feature vectors of all the training samples of each class, the EM algorithm is used to generate M models per class with different initial configurations. All of these models have a left-right topology, i.e., only looping on the same state or transitions to subsequent states are allowed. To each state, we associate m Gaussian densities. Models that have the same initial configuration are grouped together to form an individual HMM classifier.

Below is an algorithm summarising the process followed to generate the candidate classifiers. In this algorithm, the number of states is used to create diversity between the classifiers. The same principle applies for the other methods of ensemble creation, namely different initial models, different numbers of Gaussian densities and different numbers of iterations of the training algorithm.

For each $s \in S \setminus S$: the set of possible numbers of states
  For each class $\in \{c_1, c_2, \ldots, c_N\}$ the set of N classes
    Apply the EM algorithm to generate an HMM with s states.
  End
  Group established HMMs into a s-state classifier
End
Return the set of M candidate classifiers

At the end of this step, we obtain a set of M candidate classifiers (M is the cardinal of the set S representing the number of initial configurations different by their numbers of states). In this set, only K ($1 \leq K \leq M$) classifiers will be selected to be combined.

3.2.2 Selection of the best set of classifiers

One of the most important issues regarding the creation of ensembles of classifiers is the selection of the best subset. It consists in selecting adequate classifiers from a large set of different classifiers, so that the selected set can achieve optimal performance. The particularity of this selection is that it is not necessarily interested in the individual qualities of the members, but rather in the overall quality of the ensemble. The selection can be static at the training level where the selected set is used for all test examples or dynamic at the recognition level. In this later case, a new set is selected for each new test example. In this work, we use a static selection with, as selection criteria, the overall accuracy and the ensemble diversity evaluated on a validation dataset, other than that of training and test.

Note that due to the limited range of useful values of each parameter of the initial training configuration, the possible number of candidate classifiers is relatively small. Therefore, we do not use here an optimisation technique for ensembles selection, but we simply propose to randomly group the classifiers into different groups and then select the best group of classifiers among all the created groups.

The result of the training step is a set of K classifiers, where each has N models, with N being the number of classes.

3.3 Fusion and decision

Given an acoustic signal e, the N possible classes, and the ensemble of K classifiers created and selected during the training phase. To classify the new example e, it is first parameterised and represented as a sequence of feature vectors O. This operation is carried out by applying the MFCC method. Then, it passed to the set of classifiers to calculate its likelihood $P(O/\lambda_{ij})$ with respect to all the classes' models of each classifier, with λ_{ij} being the i^{th} HMM of the j^{th} class. Finally, to merge the outputs of the base classifiers, we propose to use the sum of the logarithms of likelihoods. The idea is to assign to the example to be recognised O the class $C*$ that maximises the sum of the logarithms of likelihoods given by the base classifiers for each class model, according to the following equations:

$$S_j = \sum_{i=1}^{K} \log P\left(O \setminus \lambda_i^j\right); 1 \leq j \leq N \tag{9}$$

$$C^* = argmax(S_j), 1 \leq j \leq N \tag{10}$$

4 Experimental results and discussion

- Goals. In this section, we aim to explore the following points:

  The impact of the different parameters of the initial configuration of the training algorithm on the creation of HMM-based ensembles, in terms of accuracy and diversity. For this, four parameters will be studied: The number of HMM states, the initial model of the EM algorithm, the number of Gaussian densities per state, and the number of iterations of the training algorithm.

  Robustness to intra-class variability (interspeaker variability in particular).

  The impact of each proposed ensemble creation method on 10 diversity measures taken from the literature, as well as the relationship between these measures.

  The impact of the ensemble size on accuracy.

  The relationship between diversity and overall accuracy.

  The impact of diversity on the combination gain, which is the difference between the recognition rate of the ensemble and the best recognition rate of the individual classifiers.

- Initial setting: The ranges of values and the initial configuration of the parameters, for each method of ensemble creation, are presented in Table 1. The first column of the table represents the proposed creation methods. In the second column, we indicate the number of generated candidate classifiers. The third column represents the range of values of the modified parameter that we restrict to those that are plausible. Columns 4, 5, 6 and 7 represent the chosen value respectively for the number of states, the initial model, the number of Gaussian densities per state and the number of iterations of the training algorithm.

- Diversity measures: Diversity measures allow quantifying the complementarity of individual classifiers. They can be used to study the relationship between diversity and accuracy of an ensemble of classifiers or as an ensemble selection criterion. To quantify the ensemble diversity, we used four pairwise measures and six non-pairwise measures. Where a pairwise measurement is calculated by averaging the measured values for all pairs of base classifiers, while a non-pairwise measurement is taken across all classifier outputs. Table 2 summarises the used measures and their characteristics.

Note that the diversity measures, as shown in Table 2, can be classified into two categories: The first one includes measures that reflect similarity, i.e., the lower the value (↓), the greater the diversity. The second category includes measures that reflect diversity, i.e., the higher the value (↑) the greater the diversity. Q, ρ, DF, k and θ belong to the first category, while the other measures belong to the second. A good presentation of these ten measures can be found in Kuncheva and Whitaker (2001, 2003), Shipp and Kuncheva (2002).

Table 1 Range of values and initial configuration of the training algorithm parameters for each method of ensemble creation

	Classifiers number	*Range of plausible values of the modified parameter*	*The number of states*	*The initial model*	*The number of Gaussian densities*	*The number of iterations*
Different numbers of states	26	From 5 to 30 states	/	A single random model	1	25
Different initial models	66	Random models	10	/	1	25
Different numbers of Gaussian densities	12	From 1 to 12 Gaussians	10	A single random model	/	25
Different numbers of iterations	20	From 15 to 110 with a step of 5 iterations	10	A single random model	1	/

Table 2 Diversity measures used

Diversity measurement	*Reference*	*Type*	*Similarity (↓) or diversity (↑)*
Q-statistic (Q)	Yule (1900)	Pairwise	↓
Correlation (ρ)	Sneath and Sokal (1973)	Pairwise	↓
Disagreement (D)	Ho (1998), Skalak (1996)	Pairwise	↑
Double fault (DF)	Giacinto and Roli (2001)	Pairwise	↓
Entropy of the votes (Ent)	Cunningham and Carney (2000)	*Non-pairwise*	↑
Difficulty index (θ)	Hansen Salamon (1990)	*Non-pairwise*	↓
Kohavi-Wolpert variance (kw)	Kohavi and Wolpert (1996)	*Non-pairwise*	↑
Interrater agreement (k)	Dietterich (2000), Fleiss et al. (2013)	*Non-pairwise*	↓
Generalised Diversity (GD)	Partridge and Krzanowski, (1997)	*Non-pairwise*	↑
Coincident Failure Diversity (CFD)	Partridge and Krzanowski (1997)	*Non-pairwise*	↑

- The dataset used: In order to analyse and evaluate the performance of the proposed approach, several experiments were carried out on the spoken Arabic digits (SAD) dataset (Bedda and Hammami, 2010). This is the most cited dataset for Arabic speech recognition systems in the last decade. It contains 8,800 samples uttered by 88 Arabic speakers (44 men and 44 women). The same speaker repeats each digit 10 times. Each sample is represented by a series of 13 MFCC coefficients. The training

set contains 6,600 samples spoken by 66 speakers, and the test set contains 2,200 samples spoken by 22 speakers who did not participate in the training set. The systems designed are, then, speaker-independent systems. We have divided the training dataset into two parts. The first part, containing 5,280 samples, reserved for model training, and the second, containing 1,320 samples, is used for validation. The latter is used to evaluate the quality of the models during the selection step in the training phase. Whereas, the test set is used to evaluate the final system.

We chose to use the SAD dataset for two main reasons: First, it is freely available on the Net, which allows us making a direct comparison with previous work. Second, a relatively large number of speakers participated in the creation of the dataset. This allows effectively studying the problem of interspeaker variability. However, its main limitation is the unavailability of audio files; only the MFCC features are distributed, which prevented us from testing other acoustic analysis techniques such as LPC, LPCC, PLP, etc.

Spoken digit recognition is needed in many digit-based applications, such as voice dialer, airline reservations, banking systems, forms automation, and various other areas. Spoken digits Recognition is one of the most difficult tasks in the field of speech recognition (Saleh and Wazir, 2018), especially for poorly endowed languages. Indeed, several recent works have been applied on SAD dataset, we cite, as examples, (Kamura et al., 2022; Iwana et al., 2020; Wazir and Chuah, 2019; Guerid and Houacine, 2019; Saleh and Wazir, 2018; Guerid et al., 2018; Touazi and Debyeche, 2017).

4.1 Experiment 1: Performance evaluation of the proposed approach

This experiment is carried out in order to compare the performances of the proposed approach and those of the base classifiers in terms of accuracy. It allows validating and showing the role of the complementarity of the classifiers generated from different initial configurations within the framework of a homogeneous ensemble of Markovian classifiers. The results presented in Table 3 correspond to the best-performing subset among a large number of candidate classifiers for each method of ensemble creation. The first column represents the different proposed ensemble creation methods. The second column gives the performance of the best individual classifier. The third column represents the performance of the ensemble using the majority vote rule as a merging strategy, and the last column shows the performance of the ensemble using the sum of log-likelihood rule.

Table 3 reveals that the fusion of the classifiers always gives better results, and this, compared to all the individual classifiers. This shows that exploiting the sensitivity of HMMs to the initial training configuration has a strong impact on the creation of HMM-based ensembles, mainly when using different initial models with the sum-of-likelihood rule, where we get the best accuracy (97%). A second remark can be made from Table 3 is that the sum-of-likelihoods rule exceeds the majority vote rule in all cases. Therefore, this will be used, without mentioning it, as a fusion method in all future experiments.

Table 3 Accuracy (%) of the best individual classifier and the ensemble for the four methods of ensemble creation

	Best individual accuracy	*Ensemble accuracy with majority voting rule*	*Ensemble accuracy with the sum of likelihood rule*
Different numbers of states	93.409	94.500	96.54
Different initial models	*92.854*	*95.693*	*97.000*
Different numbers of Gaussian densities	92.590	94.000	94.818
Different numbers of iterations	92.757	95.700	96.31

4.2 Experiment 2: Evaluation of robustness to data variability

We are interested here in the study of the gain in terms of robustness and stability of our approach in the face of interspeaker variability. The SAD dataset is dedicated to speaker-independent systems, because the 22 test speakers did not participate in the recording of the training dataset, which implies a significant intra-class variability due to the large interspeaker variability. A robust system should not be too sensitive to this variability. To study the robustness of our approach to interspeaker variability, we propose to use two dispersion parameters taken from probability theory and statistics, namely the standard deviation and the coefficient of variation. The lower the values of these parameters, the more robust the system, and conversely, the higher they are, the more there will be of chances that the performance of the system deteriorates strongly due to interspeaker variability.

We calculated the standard deviation of speaker accuracies. The lower this standard deviation, the more robust the system and, therefore, less sensitive to speaker variation.

The interspeaker standard deviation is calculated as follows:

$$\sigma = \sqrt{\frac{1}{n}\sum_{speak=1}^{n}\left(Acc_{s\,peak} - \mu\right)^2} \tag{11}$$

where, n is the number of speakers (here, $n = 22$), Acc_{speak} is the accuracy over the set of examples of speaker speak and μ is the average accuracy of the speakers (corresponds to the overall accuracy of the ensemble).

We also calculated the coefficient of variation (CV), known as the relative standard deviation RSD (Relative Standard Deviation). It is a relative measure of the data dispersion around the mean. This coefficient is defined as the ratio of the standard deviation σ to the mean μ, and is often expressed as a percentage (see equation 12). Its advantage is that it allows comparing variation degrees, even if the means are different. The lower the value of the coefficient of variation, the lower the dispersion around the average accuracy, and therefore, the greater the robustness of the system in the face of interspeaker variability.

$$CV = \frac{\sigma}{\mu} * 100 \tag{12}$$

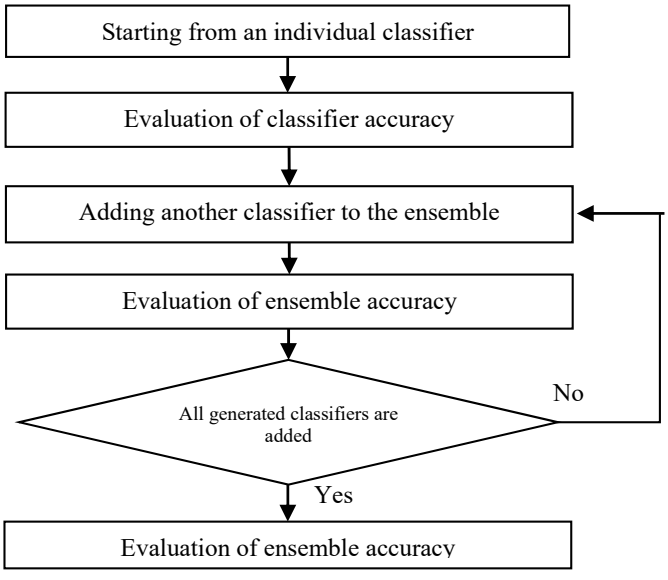
In Table 4, we summarise the results obtained in terms of robustness measured by the standard deviation and the interspeaker CV.

Table 4 Evaluation of robustness to interspeaker variability

	Basic HMM	*Ensemble with different number of states*	*Ensemble with different initial models*	*Ensemble with different numbers of Gaussians*	*Ensemble with different number of iterations*
Average Accuracy (%)	93.81	*96.54*	*97*	*94.45*	*96.31*
Ensemble size	1	4	9	3	4
σ	11.2512	*7.8177*	*5.1824*	*8.6008*	*6.6288*
CV	11.99	*8.10*	*5.34*	*9.11*	*6.88*

From Table 4, we can clearly notice the stability of our approach and its robustness to intra-class variability, especially in the case of multiple modelling from different initial models. In this case, we obtained a standard deviation of 5.1824 and a coefficient of variation of 5.34, which are much lower than those of the base HMM classifier, for which a standard deviation of 11.2512 and a coefficient of variation of 11.99 were marked.

Figure 3 Experimental protocol to study the impact of ensemble size on accuracy

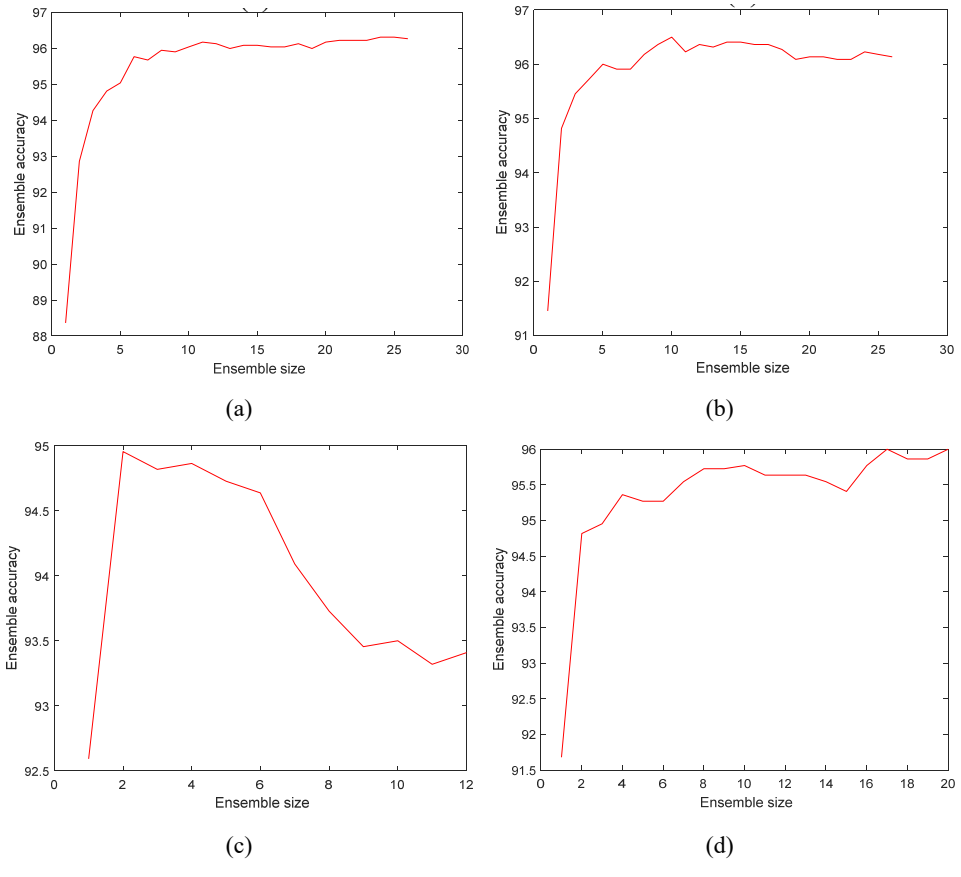


4.3 Experiment 3: The Impact of the ensemble size on accuracy

The goal of this experiment is to examine the relationship between ensemble size and accuracy for each method of ensemble creation. The experimental protocol (Figure 3) followed to achieve this objective is as follows: We start by using a single classifier, then

we gradually add another, and we evaluate the accuracy of the whole, and so on, until that all generated classifiers are used. The curves in Figure 4 show the results obtained.

Figure 4 The impact of ensemble size on accuracy in the case of: (a) different numbers of states, (b) different initial models, (c) different numbers of Gaussian densities, and (d) different numbers of iterations (see online version for colours)



An analysis of the results presented in Figure 4 leads to the following observations:

- The ensemble always gives better results than any individual classifier. This confirms the results obtained in the previous experiment.
- Ensemble size has a small impact on accuracy in all three cases: Different numbers of states (Figure 4(a)), different initial models (Figure 4(b)), and different numbers of iterations (Figure 4(d)).
- Increasing the number of classifiers does not systematically improve the accuracy of the ensemble, especially in the case of different numbers of Gaussian densities (Figure 4(c)). For this specific case, we marked an opposite impact, which can be explained by the fact that the classifier added each time is weaker than the previous classifiers. Indeed, the accuracy of the classifiers generated in this case decreases when the number of Gaussian densities increases. For example, the accuracies of the

first three classifiers are 92.59%, 92.50 and 90.59% respectively; therefore adding the third classifier to the ensemble of the first two classifiers will affect negatively and significantly the overall accuracy of the ensemble.

From these observations, we can conclude that the extension of the ensemble by the addition of a new member is not necessarily useful. It depends on how it performs relative to other members of the ensemble and its effect on diversity. Nevertheless, the size of the ensemble could have a big impact on the performance of the ensemble if the base classifiers are both high accurate and diverse.

4.4 Experiment 4: Comparison of the 4 methods of ensemble creation in terms of accuracy and diversity

In Table 5, values are calculated for 10 different randomly chosen ensembles, each of them containing 5 base classifiers. The first row represents the proposed ensemble creation methods. The second row represents the average accuracy of the 10 ensembles on validation dataset, whereas the third one reports the average accuracy of the 10 ensembles on the test dataset. The other rows indicate the average values of the different diversity measures (Q, ρ, D, DF, kw, k, Ent, θ, GD and CFD) calculated on the validation dataset. The best accuracy and diversity values are marked in bold and underlined.

Table 5 Comparison of the four methods of ensemble creation in terms of accuracy and diversity

		Different numbers of states	*Different initial models*	*Different numbers of Gaussian densities*	*Different numbers of iterations*
Average accuracy of the 10 ensembles on validation dataset		93.636	*93.846*	90.871	93.125
Average accuracy of the 10 ensembles on test dataset		95.509	*95.660*	92.977	95.306
Pairwise measures calculated on validation dataset	$Q \downarrow$	0.885	*0.856*	0.879	0.870
	$\rho \downarrow$	0.448	*0.409*	0.468	0.430
	$D \uparrow$	0.103	0.114	*0.123*	0.110
	$DF \downarrow$	0.052	*0.049*	0.073	0.052
Non-Pairwise measures calculated on validation dataset	$kw \uparrow$	0.041	0.045	*0.049*	0.044
	$k \downarrow$	0.446	*0.403*	0.465	0.425
	$Ent \uparrow$	0.150	0.165	*0.180*	0.161
	$\Theta \downarrow$	0.052	*0.049*	0.066	0.052
	$GD \uparrow$	0.495	*0.533*	0.462	0.512
	$CFD \uparrow$	0.703	*0.729*	0.677	0.711

The following conclusions can be drawn from Table 5:

Whether for the validation dataset (row 2) or for the test dataset (row 3), the four proposed methods can be ranked, in terms of accuracy, from best to worst as follows: Different initial models (93.84% and 95.66%), different numbers of states (93.63% and

95.50%), different numbers of iterations (93.12% and 95.30%) and different numbers of Gaussian densities (90.87% and 92.97%).

The majority of diversity measures (Q, ρ, DF, kw, k, Ent, θ, GD and CFD) show that the classifiers created from different initial models are the most diversified. Therefore, it can be inferred that there is an agreement between accuracy and these diversity measures.

We can clearly observe that D, kw and Ent are not suitable for predicting which of the proposed ensemble creation methods is better. Indeed, the best values of these three measures are obtained when using different numbers of Gaussian densities. However, this method is the worst in terms of overall accuracy (92.977% vs. 95.509%, 95.660% and 95.306%). Moreover, by excluding this case from our comparison, we obtain a total consensus between the ten measures of diversity.

4.5 Experiment 5: Ensemble selection

There is currently no consensus in the literature regarding the choice of a diversity measure or another measure for ensemble selection, as this essentially depends on the problem under study. This experiment aims to study the possible relationships between the different measures of diversity and the ensemble accuracy with the aim of choosing, among these measures, the one best suited to our approach, and to show if the diversity within the base classifiers is sufficient for ensemble selection. To achieve this goal, we considered five ensembles numbered from 1 to 5 for each ensemble creation method. The ten diversity measures are calculated on the validation dataset, while the accuracies of the ensembles are evaluated on the test dataset. Table 6 presents the obtained results in the case of using classifiers, which differ, in their number of states. Table 7 presents the results of using classifiers with different initial models. Table 8 presents the results obtained using classifiers with different numbers of Gaussian densities. Whereas, the results, in the case of the use of classifiers with different numbers of iterations, are reported in Table 9. In each of these tables, the first column represents the ensemble number, the second one represents the ensemble accuracy, and the other columns represent the diversity measures. For better readability, the best values are underlined.

Table 6 Relationship between diversity measures and ensemble accuracy when using different numbers of states

Ensemble number	*Ensemble accuracy (%) on the test dataset*	*Pairwise measures on validation dataset*				*Non-Pairwise measures on validation dataset*					
		Q↓	*ρ*↓	*D*↑	*DF*↓	*kw*↑	*K*↓	*Ent*↑	*Θ*↓	*GD*↑	*CFD*↑
1	95.045	*0.864*	0.401	*0.104*	0.044	*0.042*	0.398	*0.150*	0.045	0.543	0.753
2	95.636	0.902	0.424	0.081	0.035	0.032	0.421	0.115	0.037	0.534	*0.757*
3	95.727	0.879	*0.394*	0.087	*0.034*	0.035	*0.393*	0.127	*0.037*	*0.559*	0.755
4	95.227	0.918	0.460	0.074	0.037	0.029	0.458	0.108	0.039	0.501	0.708
5	95.909	0.904	0.428	0.079	0.035	0.031	0.426	0.115	0.037	0.530	0.726

Table 7 Relationship between diversity measures and ensemble accuracy when using different initial models

Ensemble number	*Ensemble accuracy (%) on the test dataset*	*Pairwise measures on validation dataset*				*Non-Pairwise measures on validation dataset*					
		Q↓	*ρ*↓	*D*↑	*DF*↓	*kw*↑	*K*↓	*Ent*↑	*Θ*↓	*GD*↑	*CFD*↑
1	96.000	0.811	0.335	*0.114*	0.034	*0.045*	0.313	*0.160*	0.037	0.623	0.811
2	96.181	0.859	0.372	0.091	0.033	0.036	0.374	0.132	0.036	0.576	0.763
3	95.545	0.864	0.371	0.090	0.032	0.036	0.369	0.130	0.035	0.580	0.770
4	*96.636*	*0.799*	*0.302*	0.107	*0.029*	0.043	*0.296*	0.155	*0.033*	*0.645*	*0.810*
5	94.772	0.888	0.425	0.090	0.040	0.036	0.423	0.130	0.042	0.526	0.731

Table 8 Relationship between diversity measures and ensemble accuracy when using different numbers of Gaussian densities

Ensemble number	*Ensemble accuracy (%) on the test dataset*	*Pairwise measures on validation dataset*				*Non-Pairwise measures on validation dataset*					
		Q↓	*ρ*↓	*D*↑	*DF*↓	*kw*↑	*K*↓	*Ent*↑	*Θ*↓	*GD*↑	*CFD*↑
1	94.727	0.887	0.426	0.094	0.043	0.037	0.425	0.136	*0.044*	0.522	0.731
2	91.227	0.907	0.508	*0.106*	0.070	*0.042*	0.506	0.157	0.065	0.432	0.646
3	*94.818*	*0.883*	*0.401*	0.089	*0.036*	0.029	*0.400*	0.134	*0.044*	*0.551*	*0.733*
4	92.454	0.916	0.503	0.093	0.058	0.031	0.502	0.140	0.063	0.445	0.647
5	91.090	0.902	0.498	0.107	0.068	0.036	0.497	*0.161*	0.071	0.441	0.628

Table 9 Relationship between diversity measures and ensemble performance when using different numbers of iterations

Ensemble number	*Ensemble accuracy (%) on the test dataset*	*Pairwise measures on validation dataset*				*Non-Pairwise measures on validation dataset*					
		Q↓	*ρ*↓	*D*↑	*DF*↓	*kw*↑	*K*↓	*Ent*↑	*Θ*↓	*GD*↑	*CFD*↑
1	95.727	0.893	0.419	0.090	0.038	0.036	0.412	*0.131*	0.040	0.538	0.743
2	94.590	0.890	0.421	*0.093*	0.040	0.037	0.411	0.133	0.041	0.537	0.747
3	*96.000*	*0.886*	*0.391*	0.080	*0.030*	*0.032*	*0.390*	0.115	*0.033*	*0.566*	*0.760*
4	95.727	0.887	0.401	0.085	0.033	0.034	0.394	0.121	0.036	0.558	0.771
5	95.045	0.908	0.445	0.082	0.038	0.032	0.440	0.120	0.040	0.514	0.712

As shown by the results presented in Table 6, there is no well-established relationship between ensemble accuracy and diversity measures in the case of different numbers of states. Indeed, the most diverse ensemble (number 3) is not the best accurate (number 5). Therefore, it is not useful to use these diversity measures as a criterion for ensemble selection.

Contrary to Table 6, the results of Tables 7, 8 and 9 clearly indicate that:

- When different initial models are used to create the ensemble (Table 7), if we exclude the measures *D*, *kw* and *Ent,* the most diverse ensemble (number 4) considering all the other diversity measures is indeed the most efficient. Similarly, for the case where different numbers of Gaussian densities are used (Table 8), the most diversified set (number 3) is also the most accurate.
- In case of using different numbers of iterations as ensemble creation method (Table 9), except for *D* and *Ent*, the most diverse ensemble in terms of all diversity measures is also the more accurate (number 3).

From the last four tables, the relationship between the different diversity measures can be summarised in Table 10. The crosses indicate the diversity measures that succeeded in selecting the best ensemble in terms of accuracy.

Table 10 Summary table of the relationship between the different diversity measures and their impact on the ensemble selection

	Q	*ρ*	*D*	*DF*	*kw*	*k*	*Ent*	*θ*	*GD*	*CFD*
Different numbers of states										
Different initial models	X	X		X		X		X	X	X
Different numbers of Gaussian densities	X	X		X		X		X	X	X
Different numbers of iterations	X	X		X	X	X		X	X	X

To conclude this experiment, as shown in Table 10, we can ensure that all diversity measures except *D*, *kw* and *Ent* can play an important role in ensemble selection, except in the case of using different numbers of states, where the most diverse ensemble is not necessarily the most accurate. Therefore, to select the best ensemble, it is enough to use one of the seven measurements *Q*, *ρ*, *DF*, *k*, *θ*, *GD*, and *CFD* because a total agreement is marked between all these measurements. If there was no total consensus, it would be possible to have them voted on to further improve the result.

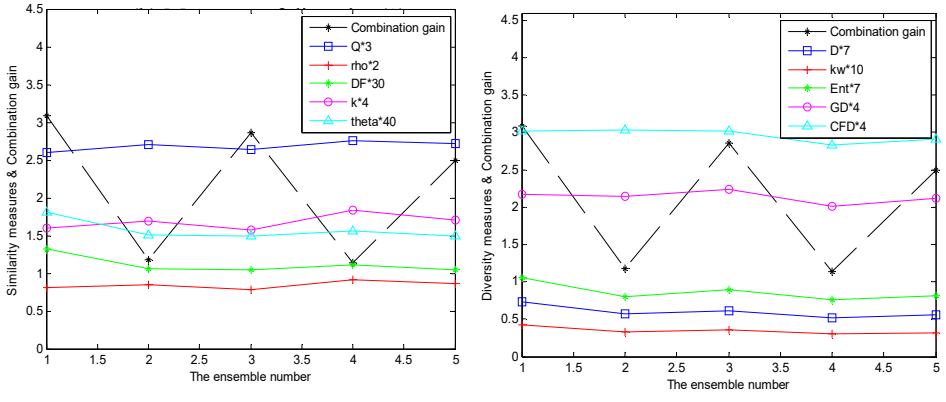
4.6 Experiment 6: The impact of diversity on combination gain

We aim through this experiment to examine the relationship between different diversity measures and the combination gain. The latter corresponds to the difference between the accuracy of the ensemble and that of the best base classifier. For this purpose, we have built several ensembles by each of the four proposed methods. For each ensemble, we calculated the combination gain and the diversity of its base classifiers using the ten diversity measures. The obtained results are illustrated in Figures 5, 6, 7 and 8, where the x-axis represents the numbers of the generated ensembles and the y-axis represents the combination gain and the diversity measures.

We have divided the diversity measures into two groups: Those that reflect the similarity between the base classifiers (*Q*, *ρ* (marked by *rho* in the figures), *DF*, *k* and *θ* (marked by theta in the figures)), they are illustrated in left Figures 5(a), 6(a), 7(a) and 8(a), and those reflecting the diversity of base classifiers (*D*, *kw*, *Ent*, *GD* and *CFD*)

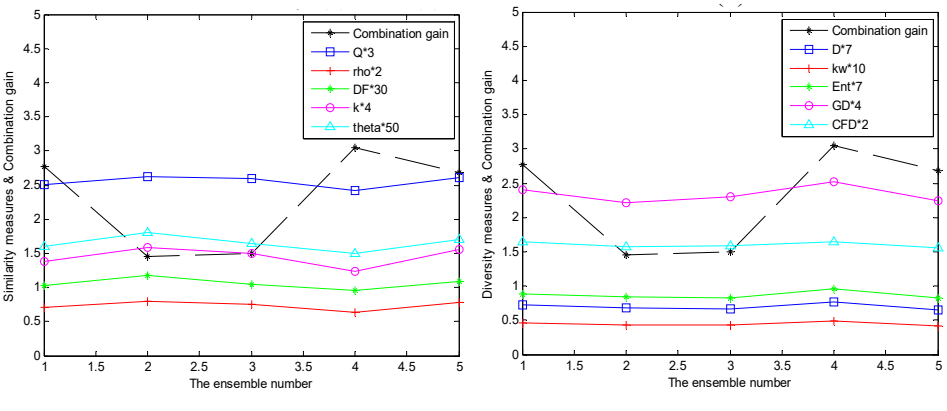
which are illustrated in right Figures 5(b), 6(b), 7(b), and 8(b). As noted earlier in this section, in the case of similarity measures, the smaller the value, the greater the diversity, and in the case of diversity metrics, the greater the value, the greater the diversity. In order to make it possible to represent several different measures in the same graph, we have multiplied each diversity measure by a positive value; this will have no effect on the relationship between the curves because all measures of the same curve are multiplied by the same value.

Figure 5 Relationship between diversity and combination gain in the case of different numbers of states: (a) measures of similarity (↓), and (b) measures of diversity (↑) (see online version for colours)



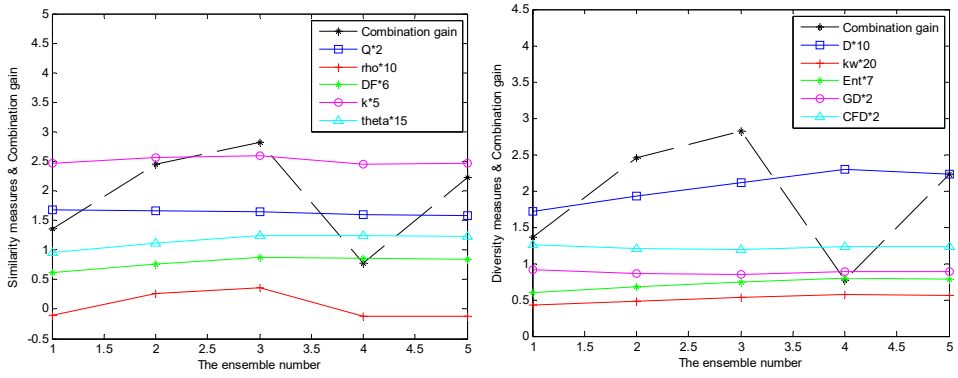
(a) (b)

Figure 6 Relationship between diversity and combination gain in the case of different initial models: (a) measures of similarity (↓), and (b) measures of diversity (↑) (see online version for colours)



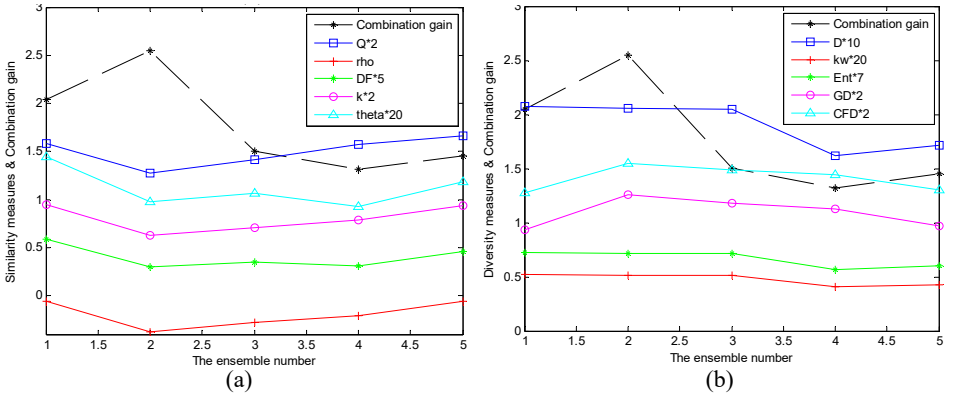
(a) (b)

Figure 7 Relationship between diversity and combining gain in the case of different numbers of Gaussian densities: (a) measures of similarity (↓), and (b) measures of diversity (↑) (see online version for colours)



(a) (b)

Figure 8 Relationship between diversity and combination gain in the case of different numbers of iterations: (a) measures of similarity (↓), and (b) measures of diversity (↑) (see online version for colours)



(a) (b)

From Figure 5 where the classifiers differ in their numbers of states, we notice that all the diversity measures except *DF* and θ of the ensemble number one (Figure5(a)), agree very well with the combination gain and with the curves in Figure 5(b). For example, the most important combination gain is achieved in the ensemble number 1 and the smallest value of each similarity measure (Figure 5(a)) and the highest value of each diversity measure (Figure 5(b)) are reached in the same ensemble. Similarly, the lowest combination gain is obtained in the least diverse ensemble (number 4) in the sense of all diversity measures.

In the case of using different initial models to create the diversity between the base classifiers (Figure 6), the combination gain is fully correlated with all the diversity measures presented in the two Figures 6(a) and 6(b). The greatest combination gain is obtained in the most diverse ensemble (number 4), and the least significant gain corresponds to the least diverse ensemble (number 2).

In the case of using ensembles of classifiers with different numbers of Gaussian densities, Figure 7 shows the existence of some correlation between the diversity

measures, but there is no relevant relationship between the gain of combination and the different measures of diversity.

From Figure 8 (the case of ensembles created with different numbers of iterations), we notice that except for the case of ensemble 5, and the case of *DF* and θ in ensemble 4, the combination gain improves as diversity increases for all diversity measures. Moreover, there is full agreement between *D*, *Ent* and *kw*, and there is also a relevant relationship between these three measurements and the combination gain.

From this experiment, it can be concluded that, in almost all cases, there is a direct correlation between diversity and combination gain. However, using diversity measures alone to assess the quality of an ensemble may not be useful, as it favours the most diverse ensembles that generate an important combination gain by neglecting individual classifiers accuracies. This could lead, if the individual classifiers are weak, to poor overall accuracy, despite a significant combination gain.

Let's take an example of 4 ensembles: E1 composed of 2 weak classifiers (accuracy of 70% and 75%) but very diversified, E2 composed of 2 strong classifiers (accuracy of 80% and 85%) but not very diversified, E3 composed of 2 weak classifiers (accuracy of 70% and 75%) and little diversified, and E4 composed of 2 strong classifiers (accuracy of 80% and 85%) and very diversified. Based on the direct correlation between the diversity of the ensemble and the combination gain, we assume that the combination gain of E1 is 7%, that of E_2 is 2%, that of E_3 is 2% and that of E_4 is 6%. The overall accuracies that can be calculated by summing the accuracy of the best individual classifier and the combination gain are summarised in Table 11.

Table 11 Example showing the complementary relationship between ensemble diversity and individual performance

Ensemble	*Diversity*	*Best individual accuracy*	*Combination gain*	*Ensemble accuracy*
E1	Very strong	75% (weak)	7%	75 + 7 = 82%
E2	Weak	85% (strong)	2%	85 + 2 = 87%
E3	Weak	75% (weak)	2%	75 + 2 = 77%
E4	Strong	85% (strong)	6%	85 + 6 = 91%

We notice that E_4 gives better results than E_1 despite the latter being more diversified and presenting a higher combination gain. This is justified by the fact that the classifiers of E_1 are much weaker. For the two ensembles E_2 and E_3, which have the same degree of diversity, we notice that E_2 is more accurate because its individual classifiers are more accurate. By comparing the ensembles E_2 and E_4, which have the same individual accuracy, we find that E_4 performs better because its individual classifiers are more diverse and, therefore, exhibit a greater combination gain.

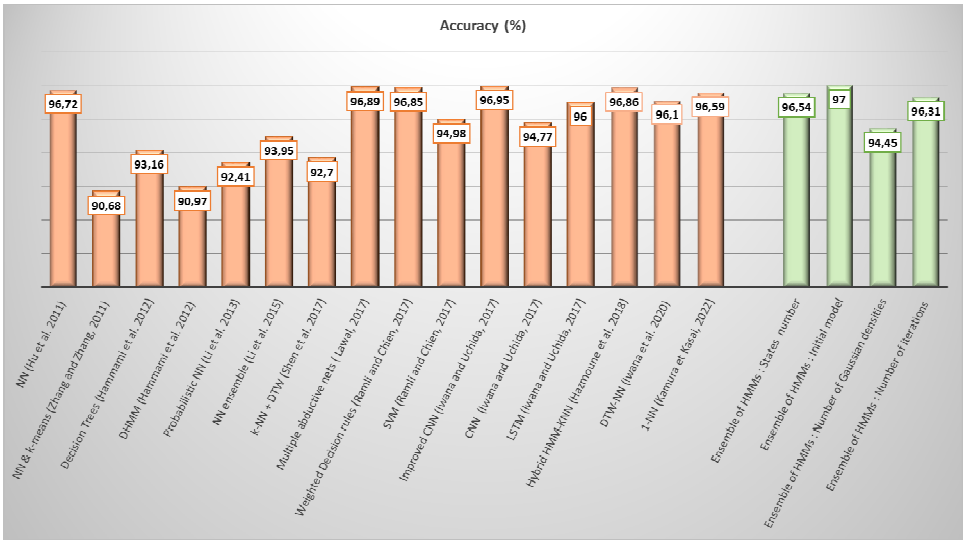
Through this example, we can deduce that there is a complementarity relationship between the ensemble diversity and the individual accuracies. To increase the accuracy of the ensemble, we must therefore select, as far as possible, the base classifiers which are both the most accurate (to obtain a high average accuracy) and the most diversified (to obtain a high combination gain).

4.7 Comparison of results with previous work on SAD dataset

To properly appreciate our ensemble approach, it is compared, as illustrated in Figure 9, with some recent works on the SAD database.

We notice from Figure 9 that the best accuracy 97% is recorded for our approach in the case of multiple modelling from different initial models. For the other cases of our approach, the accuracies are comparable with those of the best works in the literature. It is also interesting to note that, in addition to its superiority over classical approaches (basic HMM, k-NN, SVM and NN), our approach gives better results, compared to the most recent approaches based on Deep learning, such as the convolutional neural network CNN (94.77%), the improved CNN with dynamic weight alignment (96.95%), and long short-term memory (LSTM) (96%).

Figure 9 Comparison of results with previous work on SAD dataset (see online version for colours)



5 Conclusions

In this paper, we presented an HMM-based ensemble method. It consists in generating, for the same class, a set of HMM models that differ by one of the parameters of their initial configurations. The generated HMM models are, first, grouped into an initial set of classifiers that differ in their initial configurations. Then, a static selection of the best subset is performed based on accuracy and diversity as selection criteria. During the recognition phase, the selected classifiers are put in cooperation, and their outputs are merged before making the final decision. We studied four ways to differentiate between initial configurations: Different numbers of states, different initial models, different numbers of Gaussian densities per state, and different numbers of iterations of the training algorithm. The comparative results with the basic HMM and other classifiers used in previous works show the effectiveness of our approach in terms of accuracy and

robustness, especially in the case where different initial models were used as a method for ensemble creation.

An experimental analysis of the impact of these parameters on the ensemble creation and selection, in terms of diversity and accuracy, was carried out in order to determine the parameter having the greatest effect on the quality of the ensemble. We also investigated the impact of ensemble size on accuracy, as well as the relationship between ten diversity measures and the combination gain. This experimental study allowed us to draw the following conclusions. First, ensemble size has a slight impact on accuracy. Moreover, adding a new classifier to the ensemble can have an inverse impact on the quality of the ensemble. In fact, it is intuitive to argue that combining fewer high-accurate and diverse classifiers is better, in terms of performance and complexity, than combining a large number of weak classifiers. The second conclusion is that there is a strong correlation, on the one hand, between the measures Q, ρ, DF, k, θ, $GD\ et\ CFD$, and on the other hand, between D, $kw\ et\ Ent$. These last three measures are not useful for evaluating the quality of the ensemble in the four cases of ensemble creation. However, full consensus was found between all other measures Q, ρ, DF, k, θ, $GD\ et\ CFD$, where the major combination gain is recorded in the most diverse ensemble. Therefore, we can ensure that the use of one of these last measures, combined with the accuracy of the individual classifiers, as a selection criterion, will have a significant impact on the ensemble quality.

The proposed ensemble method has significant potential to be utilised in a wide range of applications where speech recognition is required, thus making the method an essential advancement in speech recognition technology. It demonstrates significant potential for further research, with several possible directions for improvement and extension. One promising avenue involves the use of optimisation techniques for ensemble selection, such as bio-inspired algorithms including gray wolf and genetic algorithms. These techniques can help identify the optimal set of base classifiers for the ensemble, leading to further improvements in accuracy. Another area for exploration is the integration of multimodal sources. By incorporating visual data like lips movements or other types of data, we can further enhance the robustness of the speech recognition system to variations in audio signals, leading to greater accuracy and reliability. In addition, the use of denoising algorithms and large noisy datasets can dramatically improve performance and robustness in noisy environments. By reducing noise levels and incorporating a broad range of noise types, the system can better distinguish speech from background noise and improve overall performance. Furthermore, combining different feature extraction methods is another fruitful strategy. By integrating a diverse range of feature sets, we can further elevate ensemble performance and capture a broader spectrum of information from the speech signals. Finally, expanding the application of this approach to other domains of pattern recognition, including but not limited to handwritten, gesture, and image recognition, and investigating its potential relevance in diverse IT domains, such as natural language processing, bioinformatics, and even anomaly detection in cyber-security, has the potential to unveil its extensive capabilities and advantages. This exploration could pave the way for fresh developments and valuable insights in these respective fields.

References

Al Dujaili, M.J., Ebrahimi-Moghadam, A. and Fatlawi, A. (2021) 'Speech emotion recognition based on SVM and KNN classifications fusion', *International Journal of Electrical and Computer Engineering,* Vol. 11, No. 2, p.1259.

Al-Hajj, R., Mokbel, C. and Likforman-Sulem, L. (2007) 'Combination of HMM-based classifiers for the recognition of Arabic handwritten words', in *Ninth International Conference on Document Analysis and Recognition*, Vol. 2, pp.959–963, IEEE.

Asafuddoula, M., Verma, B. and Zhang, M. (2017) 'An incremental ensemble classifier learning by means of a rule-based accuracy and diversity comparison', in *International Joint Conference on Neural Networks,* pp.1924–1931, IEEE.

Bakis, K. (1976) 'Continuous speech word recognition via centisecond acoustic states', in *Proc. ASA Meeting,* (Washing, DC).

Baum, L. (1972) 'An inequality and associated maximization technique in statistical estimation of probabilistic functions of a Markov process', *Inequalities*, Vol. 3, No. 1, pp.1–8.

Baum, L.E. and Eagon, J.A. (1967) 'An inequality with applications to statistical estimation for probabilistic functions of Markov processes and to a model for ecology', *Bulletin of the American Mathematical Society*, Vol. 73, No. 3, pp.360–363.

Baum, L.E. and Petrie, T. (1966) 'Statistical inference for probabilistic functions of finite state Markov chains', *The Annals of Mathematical Statistics*, Vol. 37, No. 6, pp.1554–1563.

Bedda, M. and Hammami, N. (2010) 'Spoken Arabic digit', *UCI Machine Learning Repository*, https://doi.org/10.24432/C52C9Q.

Belaïd, A. and Anigbogu, J.C. (1994) 'Mise à contribution de plusieurs classifieurs pour la reconnaissance de textes multifontes', *Traitement Du Signal*, Vol. 11, No. 1, pp.57–76.

Bhuriyakorn, P., Punyabukkana, P. and Suchato, A. (2008) 'A genetic algorithm-aided hidden markov model topology estimation for phoneme recognition of Thai continuous speech', in *2008 Ninth ACIS International Conference on Software Engineering, Artificial Intelligence, Networking, and Parallel/Distributed Computing*, August, pp.475–480, IEEE.

Biem, A. (2003) 'A model selection criterion for classification: Application to hmm topology optimization', in *Seventh International Conference on Document Analysis and Recognition*, pp.104–108, IEEE.

Bougamouza, F., Hazmoune, S. and Benmohammed, M. (2018) 'Normalisation of handwriting speed for online Arabic characters recognition', *International Journal of Computational Vision and Robotics*, Vol. 8, No. 6, pp.591–605.

Bougamouza, F., Hazmoune, S., and Benmohammed, M. (2016) 'Using Mel Frequency Cepstral Coefficient method for online Arabic characters handwriting recognition', in *5th International Conference on Multimedia Computing and Systems*, pp.87–92, IEEE.

Box, G.E. (1979) 'Robustness in the strategy of scientific model building', in *Robustness in Statistics,* pp.201–236, Academic Press, New York.

Breiman, L. (1996) 'Bagging predictors', *Machine Learning*, Vol. 24, No. 2, pp.123–140.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., ... and Amodei, D. (2020) 'Language models are few-shot learners', *Advances in Neural Information Processing Systems (NeurIPS Proceedings 2020)*, Vol. 33, pp.1877–1901.

Burges, C.J. (1998) 'A tutorial on support vector machines for pattern recognition', *Data Mining and Knowledge Discovery*, Vol. 2, No. 2, pp.121–167.

Clemente, I.A., Heckmann, M. and Wrede, B. (2012) 'Incremental word learning: efficient hmm initialization and large margin discriminative adaptation', *Speech Communication*, Vol. 54, No. 9, pp.1029–1048.

Cui, X., Zhang, W., Finkler, U., Saon, G., Picheny, M. and Kung, D. (2020) 'Distributed training of deep neural network acoustic models for automatic speech recognition: a comparison of current training strategies', *IEEE Signal Processing Magazine*, Vol. 37, No. 3, pp.39–49.

Cunningham, P. and Carney, J. (2000) 'Diversity versus quality in classification enembles based on feature selection', in *European Conference on Machine Learning,* pp.109–116, Springer, Berlin, Heidelberg.

Dempster, A. P., Laird, N.M. and Rubin, D.B. (1977) 'Maximum likelihood from incomplete data via the EM algorithm', *Journal of the Royal Statistical Society: Series B* (*Methodological*), Vol. 39, No. 1, pp.1–2.

Deng, J., Xie, X., Wang, T., Cui, M., Xue, B., Jin, Z., ... and Liu, X. (2023) 'Confidence score based speaker adaptation of conformer speech recognition systems', *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, Vol. 31, pp.1175–1190, DOI: 10.1109/TASLP.2023.3250842.

Devlin, J., Chang, M.W., Lee, K. and Toutanova, K. (2018) *Bert: Pre-Training of Deep Bidirectional Transformers for Language Understanding*. arXiv preprint arXiv:1810.04805.

Dietterich, T. G. (2000) 'An experimental comparison of three methods for constructing ensembles of decision trees: bagging, boosting, and randomization', *Machine Learning*, Vol. 40, No. 2, pp.139–157.

Ettaouil, M., Lazaar, M., and En-Naimani, Z. (2013, May) 'A hybrid ANN/HMM models for arabic speech recognition using optimal codebook', in *2013 8th International Conference on Intelligent Systems: Theories and Applications (SITA),* pp.1–5, IEEE.

Feng, W., Guan, N., Li, Y., Zhang, X. and Luo, Z. (2017) 'Audio visual speech recognition with multimodal recurrent neural networks', in *2017 International Joint Conference on Neural Networks (IJCNN)*, IEEE, May, pp.681–688.

Ferrer, M.A., Alonso, I.G. and Travieso, C.M. (2000) 'Influence of initialisation and stop criteria on HMM based recognisers', *Electronics Letters*, Vol. 36, No. 13, pp.1165–1166.

Fleiss, J.L., Levin, B. and Paik, M.C. (2013) *Statistical Methods for Rates and Proportions*, John Wiley and Sons, New York.

Forney, G.D. (1973) 'The viterbi algorithm', *Proceedings of the IEEE*, Vol. 61, No. 3, pp.268–278.

Freund, Y. (1995) 'Boosting a weak learning algorithm by majority', *Information and Computation*, Vol. 121, No. 2, pp.256–285.

Ge, Y., Zhang, X., Chen, Q. and Jiang, M. (2016) 'Initialization of the HMM-based delay model in networked control systems', *Information Sciences*, Vol. 364, pp.1–15, DOI: 10.1016/j.ins.2016.05.013.

Geng, M., Xie, X., Ye, Z., Wang, T., Li, G., Hu, S., ... and Meng, H. (2022) 'Speaker adaptation using spectro-temporal deep features for dysarthric and elderly speech recognition', *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, Vol. 30, pp.2597–2611.

Giacinto, G. and Roli, F. (2001) 'Design of effective neural network ensembles for image classification purposes', *Image and Vision Computing*, Vol. 19, Nos. 9–10, pp.699–707.

Gilpin, S.A. and Dunlavy, D.M. (2009) 'Relationships between accuracy and diversity in heterogeneous ensemble classifiers', in *SAND2009, 6940C. Department of Energy's National Nuclear Security Administration under Contract DE-AC04-94AL85000*, September.

Giuliani, D., Gerosa, M. and Brugnara, F. (2006) 'Improved automatic speech recognition through speaker normalization', *Computer Speech and Language*, Vol. 20, No. 1, pp.107–123.

Guerid, A. and Houacine, A. (2019) 'Recognition of isolated digits using DNN–HMM and harmonic noise model', *IET Signal Processing*, Vol. 13, No. 2, pp.207–214.

Guerid, A., Saboune, H. and Houacine, A. (2018) 'Recognition of isolated digits using HMM and harmonic noise model', in *2018 1st International Conference on Computer Applications and Information Security (ICCAIS)*, April, pp.1–5, IEEE.

Guo, Z., Xu, L. and Ali Asgharzadeholiaee, N. (2022) 'A homogeneous ensemble classifier for breast cancer detection using parameters tuning of MLP neural network', *Applied Artificial Intelligence*, Vol. 36, No. 1, pp.1–21.

Hamdi, A. and Frigui, H. (2015) 'Ensemble hidden Markov models with application to landmine detection', *EURASIP Journal on Advances in Signal Processing*, Vol. 2015, No. 1, p.75.

Hammami, N., Bedda, M. and Farah, N. (2012) 'Tree distributions approximation model for robust discrete speech recognition', *International Journal of Speech Technology*, Vol. 15, No. 4, pp.455–462.

Han, Y. and Boves, L. (2006) *EM Algorithm with Split and Merge in Trajectory Clustering for Automatic Speech Recognition*, Department of Language and Speech, Radboud University Nijmegen.

Hansen, L.K. and Salamon, P. (1990) 'Neural network ensembles', *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 12, No. 10, pp.993–1001.

Hazmoune, S., Bougamouza, F., Mazouzi, S. and Benmohammed, M. (2018) 'A new hybrid framework based on Hidden Markov models and K-nearest neighbors for speech recognition', *International Journal of Speech Technology*, Vol. 21, No. 3, pp.689–704.

Hazmoune, S., Bougamouza, F., Mazouzi, S. and Benmohammed, M. (2013b) 'Contributions to HMM-based speech recognition systems', *International Journal of Computational Linguistics Research*, Vol. 4, No. 1, pp.38–47.

Hazmoune. S, Bougamouza. F, Mazouzi.S, and Benmohammed. M. (2013a) 'A novel speech recognition approach based on multiple modeling by hidden Markov models', in *International Conference on Computer Applications Technology*, pp.1–6, IEEE.

He, Y., Seng, K.P. and Ang, L.M. (2023) 'Multimodal sensor-input architecture with deep learning for audio-visual speech recognition in wild', *Sensors*, Vol. 23, No. 4, p.1834.

Ho, T. (1998) 'The random subspace method for constructing decision forests', *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 20, No. 8, pp.832–844.

Hu, X., Zhan, L., Xue, Y., Zhou, W. and Zhang, L. (2011) 'Spoken arabic digits recognition based on wavelet neural networks', in *2011 IEEE International Conference on Systems, Man, and Cybernetics*, October, pp.1481–1485, IEEE.

Islam, J., Mubassira, M., Islam, M.R. and Das, A.K. (2019) 'A speech recognition system for Bengali language using recurrent neural network', in *2019 IEEE 4th International Conference on Computer and Communication Systems (ICCCS)*, February, pp.73–76, IEEE.

Itaya, Y., Zen, H., Nankaku, Y., Miyajima, C., Tokuda, K. and Kitamura, T. (2005) 'Deterministic annealing EM algorithm in acoustic modeling for speaker and speech recognition', *IEICE Transactions on Information and Systems*, Vol. 88, No. 3, pp.425–431.

Iwana, B.K. and Uchida, S. (2017) *Dynamic Weight Alignment for Convolutional Neural Networks*, arXiv: 1712.06530.

Iwana, B.K., Frinken, V. and Uchida, S. (2020) 'DTW-NN: a novel neural network for time series recognition using dynamic alignment between inputs and weights', *Knowledge-Based Systems*, Vol. 188, p.104971, https://doi.org/10.1016/j.knosys.2019.104971.

Jordan, M.I., Ghahramani, Z., Jaakkola, T.S. and Saul, L.K. (1999) 'An introduction to variational methods for graphical models', *Machine Learning*, Vol. 37, No. 2, pp.183–233.

Juang, B.H. and Rabiner, LR, (1990) 'The segmental K-means algorithm for estimating parameters of hidden Markov models', *Acoustics, Speech and Signal Processing, IEEE Transactions*, Vol. 38, No. 9, pp.1639–1641.

Kamura, M. and Kasai, H. (2022) 'A study of sequence matching method considering data transition', In 人工知能学会全国大会論文集 第, Vol. 36, No. 2022, pp.2S6IS3d01–2S6IS3d01, 一般社団法人 人工知能学会.

Khelifa, M.O., Elhadj, Y.M., Abdellah, Y. and Belkasmi, M. (2017) 'Constructing accurate and robust HMM/GMM models for an Arabic speech recognition system', *International Journal of Speech Technology*, Vol. 20, No. 4, pp.937–949.

Kim, J.H., Kim, K.K. and Suen, C.Y. (2000) 'Hybrid schemes of homogeneous and heterogeneous classifiers for cursive word recognition', *in Proc. 7th International Workshop on Frontiers in Handwriting Recognition*, pp.433–442, Amsterdam. Netherlands.

Koerich, A.L. and Poitevin, C. (2005) 'Combination of homogeneous classifiers for musical genre classification', in *IEEE International Conference on Systems, Man and Cybernetics*, Vol. 1, pp.554–559.

Kohavi, R. and Wolpert, D.H. (1996) 'Bias plus variance decomposition for zero-one loss functions', in *Proc. International Conference on Machine Learning (ICML'96)*, pp.275–283.

Kuncheva, L.I. and Whitaker, C.J. (2001) 'Ten measures of diversity in classifier ensembles: limits for two classifiers', in *A DERA/IEE Workshop on Intelligent Sensor Processing* (Ref. No. 2001/050), pp.10–11, IET.

Kuncheva, L.I. and Whitaker, C.J. (2003) 'Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy', *Machine Learning*, Vol. 51, No. 2, pp.181–207.

Kunze, J., Kirsch, L., Kurenkov, I., Krug, A., Johannsmeier, J. and Stober, S. (2017) 'Transfer learning for speech recognition on a budget', arXiv preprint arXiv:1706.00290.

Kurata, D., Nankaku, Y., Tokuda, K., Kitamura, T. and Ghahramani, Z. (2006) 'Face recognition based on separable lattice HMM', in *2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings*, Vol. 5, p.5, IEEE.

Kwong, S., Chau, C.W., Man, K.F. and Tang, K.S. (2001) 'Optimisation of HMM topology and its model parameters by genetic algorithms', *Pattern Recognition*, Vol. 34, No. 2, pp.509–522.

Lawal, I.A. (2017) 'Spoken character classification using abductive network', *International Journal of Speech Technology*, Vol. 20, No. 4, pp.881–890.

Levinson, S.E., Rabiner, L.R. and Sondhi, M.M. (1983) 'An introduction to the application of the theory of probabilistic functions of a Markov process to automatic speech recognition', *Bell System Technical Journal*, Vol. 62, No. 4, pp.1035–1074.

Li, X., Zhang, S., Li, S. and Chen, J. (2015) 'An improved method of speech recognition based on probabilistic neural network ensembles', in *2015 11th International Conference on Natural Computation (ICNC)*, August, pp.650–654, IEEE.

Li, X.G., Yao, M.F., Jian, L.R. and Li, Z.J. (2013) 'The application of probabilistic neural network in speech recognition based on partition clustering', in *Applied Mechanics and Materials*, Vol. 263, pp.2173–2178, Trans Tech Publications, https://doi.org/10.4028/www.scientific.net/amm.263-266.2173.

Lior, R. (2014) 'Data mining with decision trees theory and applications', Published by in *Series in Machine Perception and Artificial Intelligence*, 2nd ed., Vol. 81, World Scientific, Singapore, https://doi.org/10.1142/9097.

Liu, N., Davis, R.I., Lovell, B.C. and Kootsookos, P.J. (2004) 'Effect of initial HMM choices in multiple sequence training for gesture recognition', in *International Conference on Information Technology: Coding and Computing, 2004. Proceedings. ITCC 2004*, April, Vol. 1, pp.608–613, IEEE.

Looney, C.G. (1997) *Pattern Recognition Using Neural Networks: Theory and Algorithms for Engineers and Scientists*, Oxford University Press, Inc., New York.

Moghaddam, Z. and Piccardi, M. (2013) 'Training initialization of hidden Markov models in human action recognition', *IEEE Transactions on Automation Science and Engineering*, Vol. 11, No. 2, pp.394–408.

Mustafa, M.K., Allen, T. and Appiah, K. (2019) 'A comparative review of dynamic neural networks and hidden Markov model methods for mobile on-device speech recognition', *Neural Computing and Applications*, Vol. 31, No. 2, pp.891–899.

Nassif, A.B., Shahin, I., Attili, I., Azzeh, M. and Shaalan, K. (2019). Speech recognition using deep neural networks: a systematic review', *IEEE Access*, Vol. 7, pp.19143–19165, DOI: 10.1109/ACCESS.2019.2896880..

Nathan, K., Senior, A. and Subrahmonia, J. (1996) 'Initialization of hidden markov models for unconstrained on-line handwriting recognition', in: *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pp.3502–3505.

Noda, K., Yamaguchi, Y., Nakadai, K., Okuno, H.G. and Ogata, T. (2015) 'Audio-visual speech recognition using deep learning', *Applied Intelligence*, Vol. 42, No. 4, pp.722–737.

Obidziński, M. (2021) 'Response frequencies in the conjoint recognition memory task as predictors of developmental dyslexia diagnosis: a decision-trees approach', *Dyslexia*, Vol. 27, No. 1, pp.50–61.

Oruh, J., Viriri, S. and Adegun, A. (2022) 'Long short-term memory recurrent neural network for automatic speech recognition', *IEEE Access*, Vol. 10, pp.30069–30079, DOI: 0.1109/ACCESS.2022.3159339.

Palaz, D., Magimai-Doss, M. and Collobert, R. (2019) 'End-to-end acoustic modeling using convolutional neural networks for HMM-based automatic speech recognition', *Speech Communication*, Vol. 108, pp.15–32, DOI: 10.1016/j.specom.2019.01.004.

Pan, X., Chen, P., Gong, Y., Zhou, H., Wang, X. and Lin, Z. (2022) 'Leveraging unimodal self-supervised learning for multimodal audio-visual speech recognition', arXiv preprint arXiv:2203.07996.

Pardede, H.F., Adhi, P., Zilvan, V., Ramdan, A. and Krisnandi, D. (2023) 'Deep convolutional neural networks-based features for Indonesian large vocabulary speech recognition', *IAES International Journal of Artificial Intelligence*, Vol. 12, No. 2, p.610.

Partridge, D. and Krzanowski, W. (1997) 'Software diversity: practical statistics for its measurement and exploitation', *Information and Software Technology*, Vol. 39, No. 10, pp.707–717.

Qin, C.X., Qu, D. and Zhang, L.H. (2018) 'Towards end-to-end speech recognition with transfer learning', *EURASIP Journal on Audio, Speech, and Music Processing*, Vol. 2018, No. 1, pp.1–9.

Rabiner, L. R. (1989) 'A tutorial on hidden Markov models and selected applications in speech recognition', *Proceedings of the IEEE*, Vol. 77, No. 2, pp.257–286.

Rabiner, L.R. (1988) 'Mathematical foundations of hidden Markov models', in *Recent Advances in Speech Understanding and Dialog Systems,* pp.183–205, Springer, Berlin, Heidelberg.

Rabiner, L.R. and Juang, B. (1986) 'An introduction to hidden Markov models', *IEEE ASSP Magazine*, Vol. 3, No. 1, pp.4–16.

Rabiner, L.R. and Juang, B.H. (1992) 'Hidden Markov models for speech recognition–strengths and limitations', in *Speech Recognition and Understanding*, pp.3–29, Springer, Berlin, Heidelberg.

Rabiner, L.R., Levinson, S.E. and Sondhi, M.M. (1984) 'On the use of hidden Markov models for speaker-independent recognition of isolated words from a medium-size vocabulary', *AT&T Bell Laboratories Technical Journal*, Vol. 63, No. 4, pp.627–642.

Rabiner, L.R., Wilpon, J.G. and Juang, B.H. (1986) 'A segmental k-Means training procedure for connected word recognition', *AT&T Technical Journal*, Vol. 65, No. 3, pp.21–31.

Radford, A., Narasimhan, K., Salimans, T. and Sutskever, I. (2018) *Improving Language Understanding by Generative Pre-Training*, OpenIA blog [online] https://openai.com/research/language-unsupervised.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D. and Sutskever, I. (2019) 'Language models are unsupervised multitask learners', *OpenAI Blog*, Vol. 1, No. 8, p.9.

Ramli, D.A. and Chien, T.W. (2017) 'Extreme learning machine based weighting for decision rule in collaborative representation classifier', *Procedia Computer Science*, Vol. 112, pp.504–513.

Saleh, A.A. and Wazir, M.B. (2018) *Spoken Arabic Digits Recognition using Deep Learning*, PhD thesis, University of Malaya.

Samarasinghe, S. (2016) *Neural Networks for Applied Sciences and Engineering: From Fundamentals to Complex Pattern Recognition,* CRC Press, Auerbach Publications, Boca Raton, New York.

Sankar, A. (1998) 'Experiments with a Gaussian merging-splitting algorithm for HMM training for speech recognition', in *Proceedings of DARPA Speech Recognition Workshop*, February, pp.99–104.

Shen, J., Huang, W., Zhu, D. and Liang, J. (2017) 'A novel similarity measure model for multivariate time series based on LMNN and DTW', *Neural Processing Letters*, Vol. 45, No. 3, pp.925–937.

Shipp, C.A. and Kuncheva, L.I. (2002) 'Relationships between combination methods and measures of diversity in combining classifiers', *Information Fusion*, Vol. 3, No. 2, pp.135–148.

Skalak, D.B. (1996) 'The sources of increased accuracy for two proposed boosting algorithms', in *Proc. American Association for Artificial Intelligence, AAAI-96, Integrating Multiple Learned Models Workshop*, August, Vol. 1129, p.1133.

Sneath, P.H. and Sokal, R.R. (1973) *Numerical Taxonomy: The Principles and Practice of Numerical Classification*, 1st ed., WF Freeman & Co., San Francisco.

Song, Q., Sun, B. and Li, S. (2022) 'Multimodal sparse transformer network for audio-visual speech recognition', *IEEE Transactions on Neural Networks and Learning Systems*.

Sosiawan, A.Y., Nooraeni, R. and Sari, L.K. (2021) 'Implementation of using HMM-GA in time series data', *Procedia Computer Science*, Vol. 179, pp.713–720.

Spalanzani, A. (1999) *Algorithmes Évolutionnaires Pour L'étude De La Robustesse Des Systèmes De Reconnaissance De La Parole*, Doctorate dissertation, University of Joseph-Fourier-Grenoble I.

Srivastava, R.K., Shree, R., Shukla, A.K., Pandey, R.P., Shukla, V. and Pandey, D. (2022) 'A feature based classification and analysis of hidden Markov model in speech recognition', in *Cyber Intelligence and Information Retrieval*, pp.365–379, Springer, Singapore.

Ting, W. (2019) 'An acoustic recognition model for english speech based on improved HMM algorithm', in *2019 11th International Conference on Measuring Technology and Mechatronics Automation (ICMTMA)*, April, pp.729–732, IEEE.

Touazi, A. and Debyeche, M. (2017) 'An experimental framework for Arabic digits speech recognition in noisy environments', *International Journal of Speech Technology*, Vol. 20, No. 2, pp.205–224.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., ... and Polosukhin, I. (2017) 'Attention is all you need', *Advances in Neural Information Processing Systems*, Vol. 30.

Viterbi, A. (1967) 'Error bounds for convolutional codes and an asymptotically optimum decoding algorithm', *IEEE transactions on Information Theory*, Vol. 13, No. 2, pp.260–269.

Wang, D. and Zheng, T.F. (2015) 'Transfer learning for speech and language processing', in *2015 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)*, December, pp.1225–1237, IEEE.

Wazir, A.S.M.B. and Chuah, J.H. (2019) 'Spoken Arabic digits recognition using deep learning', in *2019 IEEE International Conference on Automatic Control and Intelligent Systems (I2CACIS)*, pp.339–344, IEEE.

Weston, J. and Watkins, C. (1999) 'Support vector machines for multi-class pattern recognition', in *European Symposium on Artificial Neural Networks (Esann'99)*, Bruges, Belgium, pp.219–224.

Xiao, J., Zou, L. and Li, C. (2007) 'Optimization of hidden Markov model by a genetic algorithm for web information extraction', in *International Conference on Intelligent Systems and Knowledge Engineering 2007*, Atlantis Press, pp.282–287.

Yuan, S., Zhang, J., Chen, J., Qiu, L. and Yang, W. (2019) 'A uniform initialization Gaussian mixture model–based guided wave–hidden Markov model with stable damage evaluation performance', *Structural Health Monitoring*, Vol. 18, No. 3, pp.853–868.

Yule, G.U. (1900) 'On the association of attributes in statistics: with illustrations from the material of the childhood society', *ANDC. Phil. Trans. R. Soc. Lond. A*, Vol. 194, Nos. 252–261, pp.257–319.

Zelinka, P., Sigmund, M. and Schimmel, J. (2012) 'Impact of vocal effort variability on automatic speech recognition', *Speech Communication*, Vol. 54, No. 6, pp.732–742.

Zhang, J. and Zhang, M. (2011) 'A speech recognition method based clustering neural network integration', in *2011 International Conference on Electric Information and Control Engineering*, April, pp.1120–1122, IEEE.