ACCEPTED MANUSCRIPT

Title: Problem of polish word recognition using classical speech recognition models with state allocation based on phonetic word structure.

Authors: Adrian Albrecht

To appear in: Technical Sciences

Received 16 March 2026;

Accepted 14 May 2026;

Available online 25 May 2026.

PROBLEM OF POLISH WORD RECOGNITION USING CLASSICAL SPEECH RECOGNITION MODELS WITH STATE ALLOCATION BASED ON PHONETIC WORD STRUCTURE

Adrian Albrecht
Institute of Computer Science
University of Warmia and Mazury in Olsztyn

Key words: Hidden Markov Model, phonemes, diphones, triphones, Automatic Speech Recognition, word recognition.

Abstract

Automatic speech recognition systems rely on statistical or neural models capable of modelling temporal dependencies present in acoustic signals. Among classical approaches, Hidden Markov Models (HMM) remain an important component of many speech recognition systems, particularly in tasks involving limited datasets or domain-specific vocabularies. One of the key design decisions in HMM-based systems concerns the representation of phonetic context and the number of states used to model acoustic sequences. This study investigates the impact of different phonetic representations and state allocation strategies in HMM models for the task of isolated Polish word recognition. The analysis considers three types of phonetic decomposition: phonemes, diphones and triphones. Additionally, three strategies of assigning the number of hidden states are evaluated: a constant number of states for all models, a dynamically adjusted number of states depending on the number of phonetic units in a word, and the classical speech recognition topology assuming three states per phonetic unit. Experiments were conducted on a custom dataset consisting of 3,600 recordings of 20 Polish command words spoken by nine speakers. Acoustic features were represented using MFCC coefficients and modelled with Gaussian Mixture Hidden Markov Models trained using the Baum–Welch algorithm. The obtained results indicate that dynamically assigning the number of states proportional to the number of phonemes (three states per phoneme) achieves the highest recognition accuracy. At the same time, increasing the phonetic context from phonemes to diphones and triphones did not improve performance on the analysed dataset, likely due to the increased model complexity and the limited size of the training corpus. The analysis of confusion matrices further reveals that HMM models capture phonetic similarities between words, which can lead to systematic recognition errors in phonetically similar commands.

Introduction

In recent years, artificial intelligence (AI) and machine learning techniques have significantly transformed many areas of science and technology (MAKRIDAKIS 2017). One of the domains strongly influenced by these developments is automatic speech recognition (ASR), which combines signal processing, phonetics, and computational linguistics (ŚLEDZIŃSKI 2010). Due to technological progress, speech recognition systems have become increasingly accurate and widely available, finding applications in numerous areas ranging from voice assistants such as Siri and Alexa to automatic transcription and translation systems, including speech-to-text services integrated into communication platforms.

One of the fundamental components of classical speech recognition systems are Hidden Markov Models (HMMs) (RABINER 1989), which play a significant role in modeling and analyzing sequential acoustic data (SMIT, VIRPIOJA, KURIMO 2021). Research on speech recognition using HMMs has a long history dating back to the 1970s and 1980s. Their application significantly advanced the field of speech signal processing (JELINEK 1976). It has

been demonstrated that HMMs are capable of effectively modeling temporal dependencies present in acoustic signals, which is essential for speech recognition tasks.

HMMs represent one of the most effective probabilistic approaches for modeling temporal sequences, making them well suited for speech processing, where audio signals inherently form time-dependent sequences (CHAURASIYA 2022). These models capture statistical relationships in sequential observations and enable classification of acoustic patterns corresponding to linguistic units (GALES, YOUNG 2024). In HMM-based speech recognition systems, hidden states are typically interpreted as abstract temporal stages of phonetic realization. The number of states determines the temporal granularity with which acoustic sequences are modeled. Classical ASR systems frequently employ a three-state left-to-right topology for each phonetic unit, where the states roughly correspond to the beginning, middle, and end of the phoneme articulation process. However, the optimal number of states may depend on dataset size, phonetic complexity, and variability of the acoustic material.

In recent years, deep learning approaches have achieved dominant performance in large-scale speech recognition tasks. Architectures based on deep neural networks, recurrent neural networks, and transformer models significantly outperform classical statistical approaches when trained on large speech corpora (HINTON et al. 2012; DEL-AGUA et al. 2018). Nevertheless, HMM-based systems remain valuable in research scenarios involving small datasets, domain-specific vocabularies, low computational complexity requirements, and interpretable phonetic modeling. Unlike large neural architectures, HMMs allow direct analysis of temporal state transitions and phonetic dependencies between acoustic units, making them useful analytical tools for experimental phonetic studies.

One of the important design aspects in speech recognition systems concerns the phonetic representation of words, particularly the decomposition of words into phonemes, diphones, or triphones (CHEN, MAK, LEUNG, SIVADAS 2014). Although this problem has been widely studied, language-specific characteristics remain relatively underexplored (TACHBELIE, ABATE, BESACIER 2014). While English is the most extensively researched language in speech recognition, other languages such as Polish present additional challenges due to their complex phonetic structure and numerous phonetic similarities between words, similarly to languages such as Russian (SAVCHENKO 2013). Polish belongs to the group of highly inflectional Slavic languages characterized by complex consonant clusters, rich phonetic variation, and numerous phonetic assimilation phenomena. These characteristics increase the difficulty of acoustic modeling, particularly in small-vocabulary ASR systems, where subtle phonetic similarities between words may significantly affect classification performance.

Modern deep learning ASR systems typically require extremely large speech corpora and substantial computational resources in order to achieve state-of-the-art performance. In contrast, HMM-based systems may still provide competitive and computationally efficient solutions in low-resource environments, particularly when dealing with small vocabularies and limited training datasets (BESACIER et al. 2014).

Although several works addressing Polish speech recognition exist (PONDEL-SYCZ, BILSKI 2024), limited research has investigated how different phonetic decompositions influence the modeling of relationships between phonetic components within words belonging

to a set of phonetically similar commands. This problem constitutes the main focus of the present study.

Compared to English-language ASR research, relatively few studies investigate phonetic dependency modeling in Polish speech recognition, particularly in the context of context-dependent acoustic units such as diphones and triphones. Consequently, many assumptions commonly adopted in English ASR systems may not directly generalize to Polish due to substantial phonetic and articulatory differences between the languages.

Isolated word recognition constitutes one of the classical problems in automatic speech recognition research. In contrast to continuous speech recognition, isolated word recognition assumes that individual utterances are separated in time and contain a single spoken command or word. Such systems remain relevant in embedded systems, robotics, industrial control interfaces, and assistive technologies, where low computational complexity, low latency, robustness to limited datasets, and limited vocabularies are important constraints.

Context-dependent acoustic units such as diphones and triphones are commonly used in speech recognition systems because they allow the model to capture coarticulation effects occurring between neighboring phonetic units. While phoneme-based models represent individual phonetic sounds independently, diphone and triphone representations incorporate local phonetic context, potentially improving the modeling of pronunciation variability (JURAFSKY, MARTIN 2013).

The research problem addressed in this work concerns the analysis of how the structure of hidden states in HMM models influences the modeling of phonetic dependencies between elements of Polish words. In particular, the study investigates the effect of representing words using different phonetic units—phonemes, diphones, and triphones—and the influence of the number of hidden states on the effectiveness of modeling these dependencies in the task of isolated word recognition. The main research question can therefore be formulated as follows: How do the number and allocation strategy of hidden states in HMM models influence the effectiveness of modeling phonetic units in the task of Polish word recognition?

The novelty of the present study lies not only in comparing different HMM state allocation strategies, but also in analyzing how various phonetic decompositions influence the emergence of systematic phonetic confusions between Polish command words. Particular attention is devoted to the interpretability of confusion patterns and their relationship to Polish phonetic phenomena.

The working hypothesis assumes that increasing the phonetic context of modeled units (transitioning from phonemes to diphones and triphones) may improve recognition performance, provided that the number of hidden states in the HMM model is appropriately adjusted. It is further assumed that dynamically assigning the number of states based on the number of phonetic units in a word may better reflect the phonetic structure of words compared to using a fixed number of states for all models. The objective of this study is therefore to experimentally evaluate the influence of different state allocation strategies in HMM models on the accuracy of Polish word recognition using phoneme-, diphone-, and triphone-based representations. Three modeling strategies are compared:

(1) a constant number of hidden states for all words,
(2) dynamic allocation of states depending on the number of phonetic units in a word, and
(3) the classical ASR topology assigning three states per phonetic unit.

Material and Methods

1. Dataset

The dataset (hereafter referred to as the command dataset) was created by the author of this study specifically for the purposes of training, testing, and interpreting artificial intelligence models specialized in recognizing isolated Polish words using Hidden Markov Models (HMM). The dataset is in the Polish language and contains a vocabulary consisting of 20 different Polish words representing digits and basic control commands:

- digits from zero to nine in polish,
- start and stop,
- lewo and prawo („left" and „right")
- góra and dół („up" and „down"),
- puść ("release"),
- złap ("grab"),
- oś ("axis"),
- chwytak ("gripper").

The composition of this vocabulary was inspired by an experiment conducted by Ewa Figielska in her research on evolutionary methods for training Hidden Markov Models (FIGIELSKA 2011). Each word appears 180 times in the dataset, which corresponds to 180 individual audio recordings stored in the .wav format with a sampling frequency of 44.1 kHz. Although 16 kHz sampling is commonly used in ASR systems, recordings were preserved at 44.1 kHz in order to avoid information loss during preprocessing. Each recording was manually trimmed so that it contained only the pronunciation of the corresponding word. The total number of recordings in the dataset is therefore 3,600 audio files. The dataset contains recordings from nine speakers, including the author of this study as well as eight students and graduates of the Computer Science program at the Faculty of Mathematics and Computer Science, University of Warmia and Mazury in Olsztyn. Consequently, all recordings were produced by nine different individuals, which constitutes an important characteristic of the dataset. The dataset size limits the effectiveness of higher-order phonetic context modeling.

The dataset is evenly distributed across speakers. This means that out of the 180 occurrences of each word, 20 recordings originate from each speaker. This balanced distribution plays an important role in the later stages of the analysis. The dataset is organized in a single main directory containing audio data. Inside this directory there are subdirectories corresponding to individual words included in the recognition vocabulary. Each subdirectory contains the previously mentioned .wav files, each representing a single spoken instance of the corresponding word. The audio files follow a standardized naming convention: word [X] [Y]–[Z].wav, where X represents the first 3–4 letters of the speaker's first name, Y represents the first 3–4 letters of the speaker's surname, Z denotes the sequential number of the recording

starting from 1. Example file names include „chwytak Adr Alb–1.wav", „chwytak Dom Siel–12.wav", etc..

Each speaker recorded the audio files in a designated location selected by the author that was relatively isolated from environmental noise which could potentially be interpreted by the AI algorithm as meaningful acoustic information. The selection of speakers was completely random, meaning that no specific criteria were applied during recruitment. Nevertheless, the group of participants naturally included individuals with different speaking styles and even certain speech impairments, such as rhotacism, commonly referred to as *"reranie"* in Polish, which involves difficulty in correctly pronouncing the phoneme *r*. This characteristic represents one of the many potential challenges present in the dataset.

The dataset was intentionally recorded in a manner that introduced natural variability between utterances. Speakers were instructed to pronounce each word with varying speaking rates, intonation patterns, emotional emphasis, and articulation styles across recordings. As a result, the dataset already contains a degree of natural acoustic variability that partially resembles augmentation-related effects commonly used in ASR research. However, explicit artificial augmentation techniques such as time stretching, noise injection, or pitch shifting were not applied in the present study and remain a subject for future work.

After analyzing the recorded utterances, several phonetic challenges were identified that the speech recognition model must address. For example, in the word "zero", initial devoicing may occur, meaning that the phoneme z may be pronounced as s. Additionally, the phoneme r may present difficulties for speakers affected by rhotacism and may be substituted with phonemes corresponding to l, j, or other possible realizations associated with this speech condition. In the word "dwa", progressive devoicing within the word may occur, where the phoneme w may be realized as f due to the influence of the preceding consonant. In addition, initial devoicing may also occur, causing the phoneme d to be realized as t. In the word "pięć", two phonemes that are relatively difficult to pronounce, ę and ć, appear consecutively. This may lead to the simplification of the nasal vowel ę to a phoneme corresponding to e. In the word "sześć", the phenomenon of epenthesis may occur, meaning the insertion of an additional phoneme in order to facilitate pronunciation. In this case, a phoneme corresponding to j may appear before the consonant cluster ś and ć, resulting in a pronunciation similar to "szejść". In the word "złap", initial devoicing may again occur, where the phoneme z may be realized as s due to its position at the beginning of the word. In the word "oś", devoicing may occur at the end of the word, sometimes accompanied by palatalization caused by the addition of a vowel-like sound resembling i. Other phonetic processes also occur in words included in the vocabulary. However, these are phonetic phenomena that occur consistently during the pronunciation of the given words. One example is progressive devoicing within the word "chwytak", where the phoneme w may be realized as f. Because such processes are inherent to the pronunciation of these words and occur systematically, they were not analyzed in detail, as they do not significantly affect the functioning of the recognition algorithm.

Since the recordings already contain natural variations in speaking rate and articulation dynamics, the obtained results partially reflect the robustness of the evaluated HMM configurations to speech rate variability. Nevertheless, a controlled experimental analysis

involving explicit speech-rate modification techniques remains an important direction for future work.

1. Signal Processing and Feature Extraction

To represent the speech signal, Mel-Frequency Cepstral Coefficients (MFCC) were used, which constitute one of the most widely applied acoustic representations in automatic speech recognition systems. MFCC features are commonly employed because they approximate the perceptual properties of the human auditory system through the application of the Mel frequency scale. The MFCC coefficients were computed using the librosa library in Python. The Fast Fourier Transform (FFT) was performed on analysis windows consisting of 2048 samples. Consecutive frames of the signal were shifted using a hop length of 1024 samples. For each frame, 20 MFCC coefficients were extracted. Given the sampling frequency of 44.1 kHz, the selected parameters correspond to an analysis window of approximately 46 ms, with a frame shift of approximately 23 ms between consecutive frames. Only static MFCC coefficients were used in order to focus on the influence of HMM topology rather than feature engineering. The resulting feature vectors were subsequently standardized using the StandardScaler method from the scikit-learn library. This transformation rescales the features to a distribution with a mean of 0 and a standard deviation of 1. Such normalization is commonly applied to improve the numerical stability and convergence properties of statistical learning algorithms.

2. Model Definition and Cross-Validation Procedure

For the experiment, a variant of cross-validation known as Monte Carlo Cross-Validation (also referred to as repeated random sub-sampling validation) was employed. The experiment used cross validation with 10 repetitions. In each repetition, 90% of recordings were used for training and 10% for testing. The evaluation does not enforce speaker independence, as recordings from all speakers may appear in both training and testing subsets. The models were trained using the training data, while their performance was evaluated on the testing data. In this study, Hidden Markov Models (HMMs) were utilized to model acoustic sequences corresponding to individual spoken words. Each word in the dataset was represented by a separate HMM, trained using recordings of that particular word produced by multiple speakers. The implementation employed Gaussian Mixture Hidden Markov Models (GMM-HMM) with diagonal covariance matrices. Model parameters were estimated using the Expectation–Maximization (EM) algorithm, commonly referred to as the Baum–Welch algorithm in the context of Hidden Markov Models. The classification process consisted of computing the log-likelihood of an observation sequence for each trained HMM model. The recognized word was determined by selecting the model that produced the highest log-likelihood value for the given input sequence. Depending on the experimental scenario, the number of hidden states in the HMM models was determined using different strategies. These included:

- assigning a fixed number of states for all models, or
- determining the number of states based on the number of phonetic units contained in a given word.

An overview of the applied configuration strategies is presented in Table 1.

WORD	NO. OF PHONEMES	NO. OF DIPHONES	NO. OF TRIPHONES
zero	4	3	2
jeden	5	4	3
dwa	3	2	1
trzy	3	2	1
cztery	5	4	3
pięć	5	4	3
sześć	4	3	2
siedem	5	4	3
osiem	4	3	2
dziewięć	7	6	5
start	5	4	3
stop	4	3	2
lewo	4	3	2
prawo	5	4	3
góra	4	3	2
dół	3	2	1
puść	4	3	2
złap	4	3	2
oś	2	1	1
chwytak	6	5	4

Table 1: Division of words into the phonetic part of the word

Due to the complex phonetic structure of Polish words, a simplified phonetic segmentation scheme was adopted for most words. This approach assumes that each model must contain at least one hidden state. The segmentation was based on the following relations:

$$\text{Number of diphones} = \text{number of phonemes} - 1$$

$$\text{Number of triphones} = \text{number of phonemes} - 2$$

An example of such phonetic decomposition is presented in Table 2.

WORD	PHONEMES	DIPHONES	TRIPHONES
jeden	j – e – d – e – n	je – ed – de – en	jed – ede – den

Table 2: An example of decomposing a word into phonemes, diphones and triphones

During model training, multiple observation sequences corresponding to recordings of the same word were used. These sequences were concatenated into a single observation matrix, while their individual lengths were provided to the training algorithm in the form of a length vector. This approach enables the HMM training procedure to process multiple recordings of varying duration simultaneously. For each test sample, the log-likelihood value was computed with respect to all trained HMM models. The sample was then assigned to the class corresponding to the model that produced the highest log-likelihood value.To evaluate classification performance, recognition accuracy and confusion matrices were used as the primary evaluation metrics.

Results and Discussion

After completing all experiments, the results were aggregated into a summary plot presenting the overall system accuracy as a function of the number of hidden states in each HMM model (Figure 1).

For a dataset of relatively limited size and specific structure, the best-performing approach turned out to be a dynamic state allocation strategy, where the number of states in each HMM model was determined as three times the number of phonemes in the corresponding word. This approach is consistent with commonly applied practices in automatic speech recognition systems (JURAFSKY 2013). In the conducted experiments, the triphone-based representation resulted in lower recognition accuracy compared to the diphone-based representation, which itself performed slightly worse than the phoneme-based representation. This behavior may be explained by the larger number of model parameters associated with more complex phonetic representations, combined with the limited number of available training samples. Such conditions may lead to the overfitting phenomenon, where the model captures training data patterns too specifically and fails to generalize effectively to unseen samples.

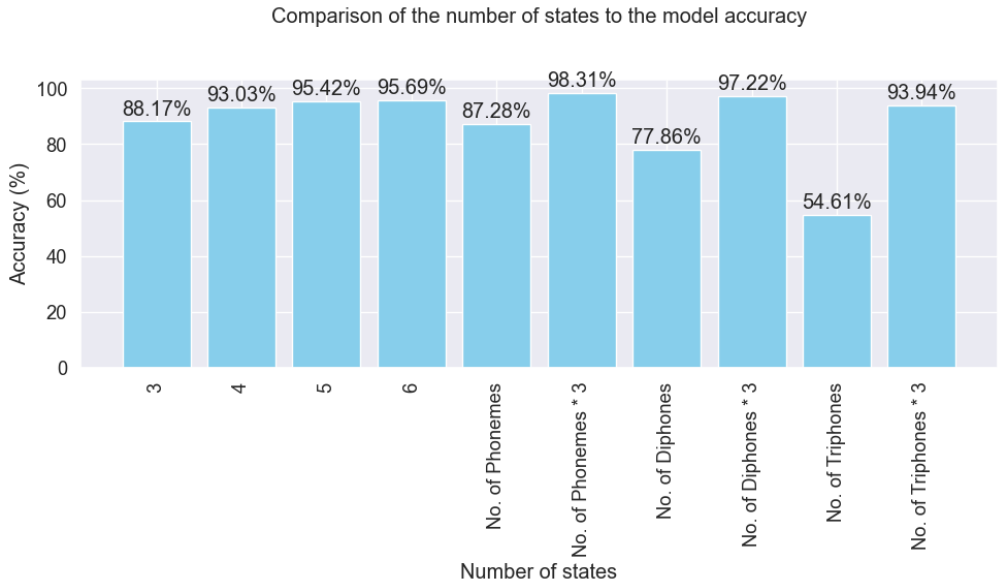

Figure 1: Comparison of the number of states to the model accuracy

It is also worth noting that models with a fixed number of hidden states, particularly those containing five or six states, achieved performance levels relatively close to the best-performing configuration. Another important aspect concerns the training time of the models. For systems with a fixed number of states, the total training time was approximately 15 minutes, whereas configurations based on traditional ASR state allocation strategies required up to 55 minutes of training. A more detailed comparison can be observed in the confusion matrices corresponding to different state allocation strategies:

- models with a fixed number of states (e.g., six states, Figure 2),
- models with the number of states equal to the number of phonemes (Figure 3),
- models with the number of states equal to three times the number of phonemes (Figure 4).

These comparisons provide additional insight into the classification behavior of the models and highlight specific patterns of misclassification between individual words.

When talking about word dependencies, it is crucial to compare the results for models that divide the number of states according to the number of phonemes (Figure 3), diphones (Figure 5) and triphones in a word (Figure 6).

At first glance, the obtained results might appear unfavorable. However, the primary objective of this study was not to confirm the superiority of the classical state allocation strategy used in automatic speech recognition systems, but rather to demonstrate the properties of Hidden Markov Models (HMMs) as statistical models capable of capturing complex phonetic relationships between words, while simultaneously revealing potential pitfalls related to phonetic similarity between them. A closer analysis of the presented results indicates that each phonetic representation captures different types of phonetic dependencies and similarities between words.

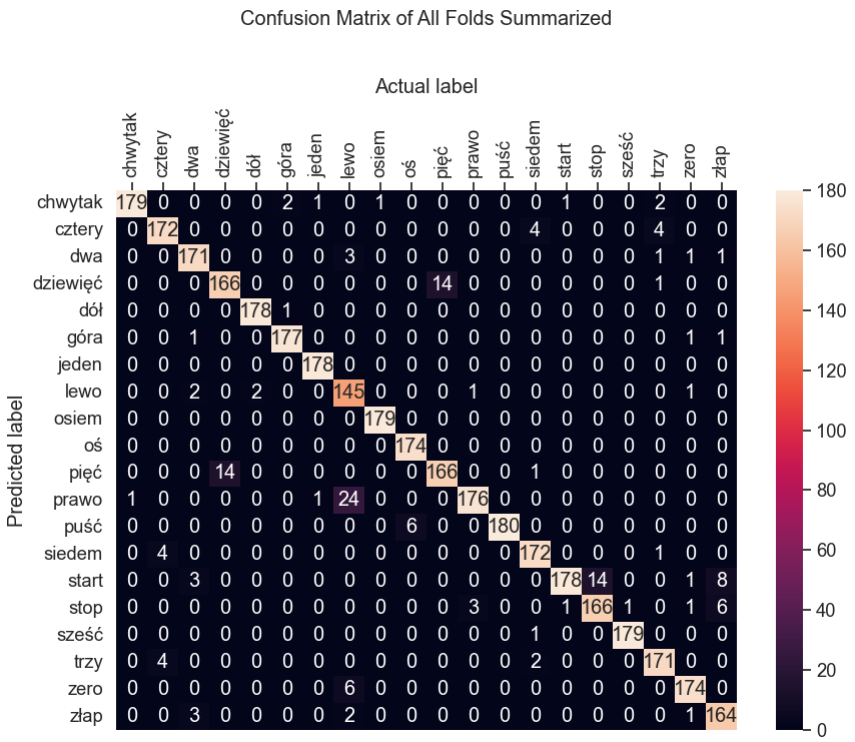

Figure 2: Confusion matrix of HMM models with a fixed number of states

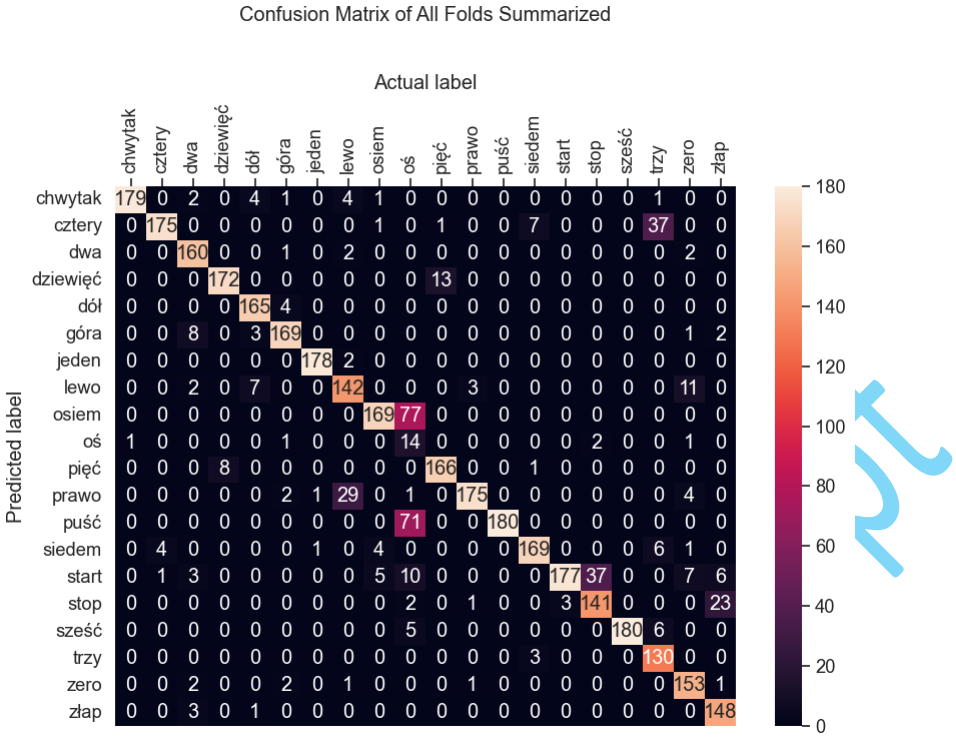

Figure 3: Confusion matrix of HMM models with the number of states equal to the number of phonemes

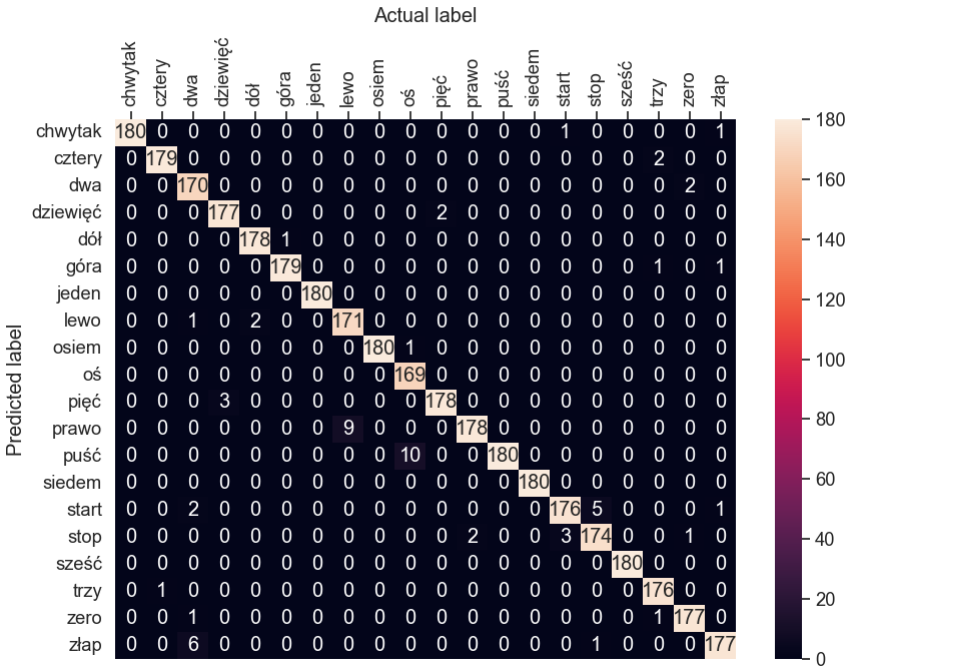

Figure 4: Confusion matrix of HMM models with the number of states equal to three times the number of phonemes

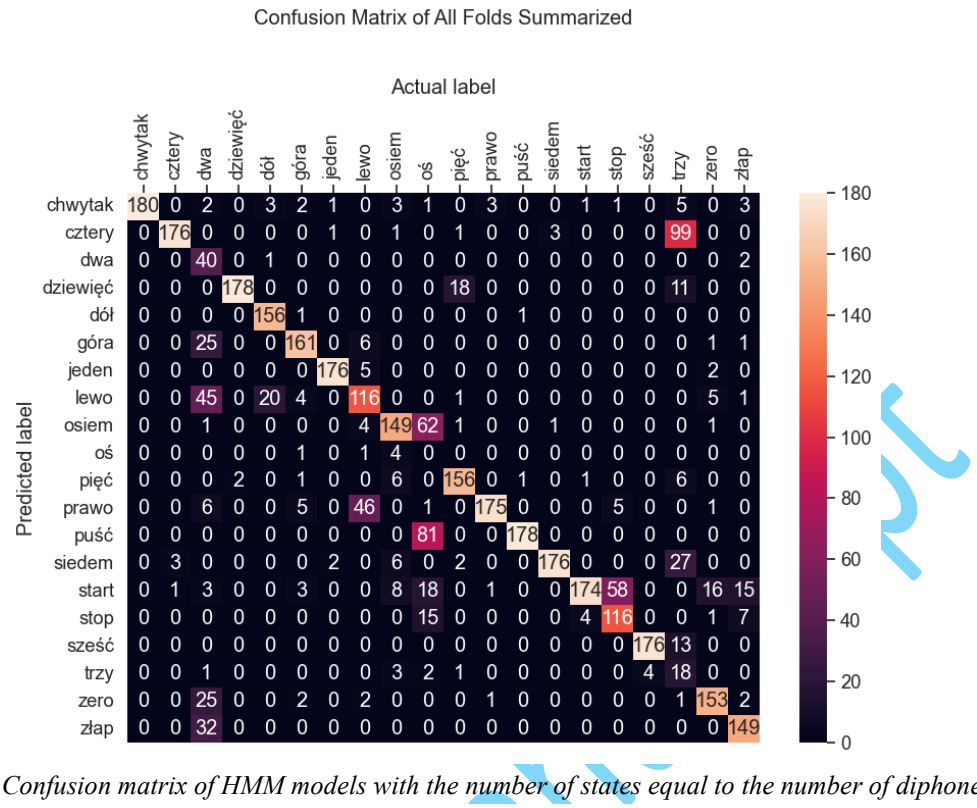

Figure 5: Confusion matrix of HMM models with the number of states equal to the number of diphones

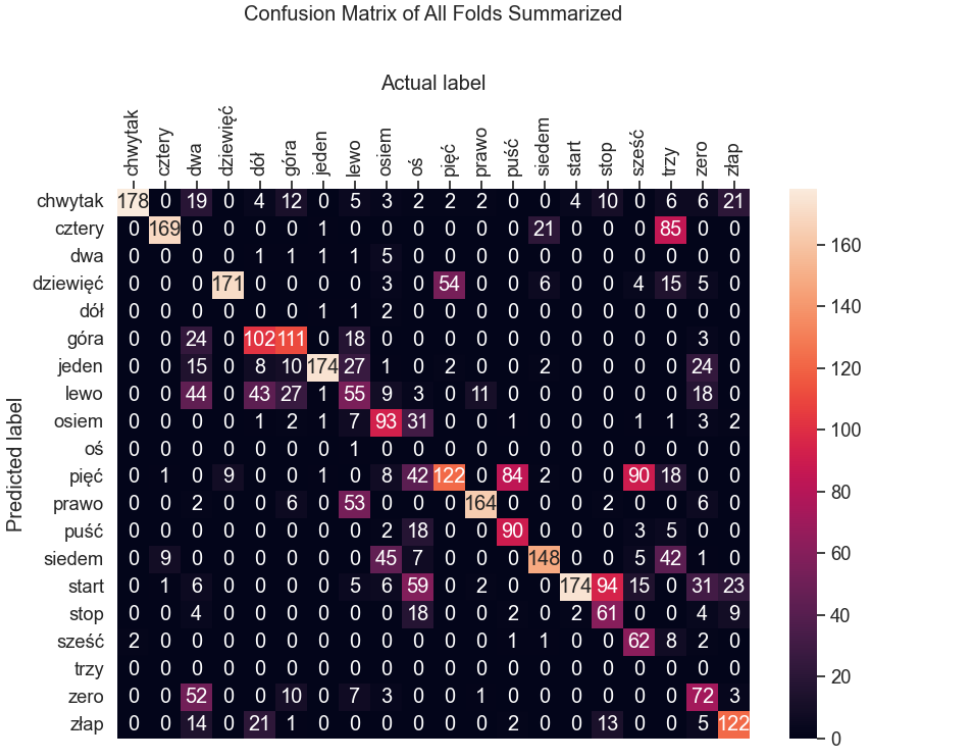

Figure 6: Confusion matrix of HMM models with the number of states equal to the number of triphones

In the phoneme-based model, it can be observed that almost no word was classified as “oś”. However, the true word “oś” was misclassified 77 times as “osiem”, likely due to the similarity of their phonetic prefixes, and 71 times as “puść”, which may be explained by the phonetic similarity of their suffixes. Similarly, due to the high phonetic similarity in their structure, the word “stop” was misclassified 37 times as “start”. A comparable relationship can be observed

in the case of 37 misclassifications of “trzy” as “cztery”. An interesting phonetic dependency is also revealed by 23 misclassifications of the word “złap” as “stop”, which may indicate that the model encountered a phonetic trap related to devoicing phenomena occurring in the pronunciation of “złap”. A similar situation occurs in the 29 misclassifications of “lewo” as “prawo”, which can be attributed to the phonetic similarity between the two words.

In the diphone-based model, several previously observed misclassification patterns became even more pronounced. In particular, the confusion between “oś” and “osiem”, “oś” and “puść”, “trzy” and “cztery”, as well as “lewo” and “prawo”, occurred more frequently. However, new phonetic relationships also emerged. The word “dwa”, which did not cause significant difficulties in the phoneme-based model, was frequently misclassified as “góra” and “złap”. This may be related to similarities in the pronunciation of diphones such as “-wa”, “-ła”, and “-ra”. Additionally, “dwa” was occasionally confused with “lewo” and “zero”. This phenomenon may be explained by more complex linguistic factors related to speech articulation errors, such as the substitution of the phoneme “-a” with “-o” following the hard consonant “w”, which may be softened or altered during pronunciation. In the case of “zero”, further phonetic variation may occur if the phoneme “r” is replaced by “l”, a phenomenon associated with certain speech articulation difficulties. Another noteworthy observation is the emergence of a weak but noticeable relationship between “pięć” and “dziewięć”, which may be related to the similar prosodic emphasis placed on the strongly articulated final diphone “-ęć”.

In the triphone-based model, some of these relationships became even more pronounced. For instance, the confusion between “pięć” and “dziewięć” increased, which may be explained by the identical accented triphone suffix “-ięć” present in both words. At the same time, several additional dependencies emerged. The word “osiem” was misclassified 45 times as “siedem”, which may be attributed to similarities between the triphone sequences “-sie” occurring in both words and the phonetic resemblance between “-iem” and “-dem”. The latter may undergo phonetic simplification or devoicing during articulation, resulting in acoustic realizations closer to “-tem”, “-eem”, or “-iem”. A similar phonetic relationship may explain the confusion between “puść” and “pięć”. The strong articulatory emphasis on consonants such as “p”, “ś”, and “ć”, combined with the nasal vowel “ę”, may contribute to pronunciation variability and increased similarity between the corresponding triphone patterns. An analogous relationship can also be observed in misclassifications between “sześć” and “pięć”.

Conclusions

This study presented an experimental analysis of the influence of state allocation strategies in Hidden Markov Models on the effectiveness of Polish spoken word recognition. Three model construction strategies were considered:

- the use of a fixed number of hidden states,
- a dynamic allocation of states based on the number of phonetic units,
- and the classical approach used in automatic speech recognition systems, where each phonetic unit is represented by three hidden states.

The experimental results indicate that, for the analyzed dataset, the best performance was achieved using a dynamic state allocation strategy, where the number of states was defined

as three times the number of phonemes in the word. This observation is consistent with traditional approaches used in automatic speech recognition (ASR) systems.

At the same time, it was observed that increasing the phonetic context by introducing diphone and triphone representations did not lead to improved classification performance. In particular, the triphone-based representation resulted in lower recognition accuracy compared to both phoneme-based and diphone-based representations. This phenomenon may be explained by the limited size of the training dataset combined with the increased number of model parameters, which can lead to the overfitting effect. Under such conditions, more complex models incorporating richer phonetic context may exhibit reduced generalization ability when applied to unseen test data.

The analysis of the confusion matrices demonstrated that HMM models are capable of capturing meaningful phonetic relationships between words, while simultaneously exhibiting susceptibility to classification errors caused by phonetic similarity between individual lexical items. The observed misclassifications indicate the influence of both shared phonetic prefixes and suffixes as well as articulatory similarities between words.

The obtained results suggest that for relatively small datasets, an optimal compromise between model complexity and generalization capability may be achieved by employing phoneme-based or diphone-based representations combined with an appropriately selected number of HMM states. Furthermore, the classical three-states-per-phoneme topology, although widely used in ASR systems, may not always be optimal for smaller datasets. In such cases, it increases both computational complexity and training time, while providing only a marginal improvement in recognition accuracy compared to models using a fixed number of states.

The presented study should therefore be interpreted primarily as an exploratory analysis of phonetic dependency modeling in HMM-based Polish ASR systems rather than as an attempt to achieve state-of-the-art recognition accuracy.

The main limitations of this study include the relatively small dataset size and the limited number of speakers. Future research should therefore involve larger speech corpora, controlled augmentation techniques, speech-rate normalization experiments, and comparisons with neural-network-based and hybrid HMM-DNN ASR architectures.

References

Besacier, L., Barnard, E., Karpov, A., Schultz, T. 2014. Automatic speech recognition for under-resourced languages: A survey. Speech Communication, 56.

Chaurasiya, H. 2022. Cognitive hexagon-controlled intelligent speech interaction system. IEEE Transactions on Cognitive and Developmental Systems, 14(4).

Chen, D., Mak, B., Leung, C.-C., Sivadas, S. 2014. Joint acoustic modeling of triphones and trigraphemes by multi-task learning deep neural networks for low-resource speech recognition. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Florence, Italy, 5592–5596.

Del-Agua, M. A., González-Domínguez, J., López-Moreno, I., Moreno, P. J. 2018. Speaker-adapted confidence measures for automatic speech recognition using deep bidirectional

recurrent neural networks. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 26(7), 1198–1206.
Figielska, E. 2011. Ewolucyjne metody uczenia ukrytych modeli Markowa. Zeszyty Naukowe Warszawskiej Wyższej Szkoły Informatyki.
Gales, M., Young, S. 2008. The application of hidden Markov models in speech recognition. Foundations and Trends in Signal Processing, 1(3), 195–304.
Hinton, G., Deng, L., Yu, D., Dahl, G., Mohamed, A., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T., Kingsbury, B. 2012. Deep neural networks for acoustic modeling in speech recognition. IEEE Signal Processing Magazine, 29(6).
Jelinek, F. 1976. Continuous speech recognition by statistical methods. Proceedings of the IEEE, 64(4).
Jurafsky, D., Martin, J. H. 2013. Speech and Language Processing. Pearson Education, Upper Saddle River.
Makridakis, S. 2017. The forthcoming Artificial Intelligence revolution: Its impact on society and firms. Futures, 90.
Pondel-Sycz, K., Bilski, P. 2024. A system dedicated to Polish automatic speech recognition – overview of solutions. Bulletin of the Polish Academy of Sciences: Technical Sciences.
Rabiner, L. R. 1989. A tutorial on hidden Markov models and selected applications in speech recognition. Proceedings of the IEEE, 77(2).
Savchenko, A. V. 2013. Phonetic words decoding software in the problem of Russian speech recognition. Automation and Remote Control, 74, 1225–1232.
Sledzinski, D. 2010. Fonemy, difony, trifony i sylaby – charakterystyka jednostek na podstawie korpusu. Kwartalnik Językoznawczy, 3–4.
Smit, P., Virpioja, S., Kurimo, M. 2021. Advances in subword-based HMM-DNN speech recognition across languages. Computer Speech and Language, 66, 101158.
Tachbelie, M. Y., Abate, S. T., Besacier, L. 2014. Using different acoustic, lexical and language modeling units for automatic speech recognition of an under-resourced language – Amharic. Speech Communication, 56.