UDC 004.032.26
DOI: 10.15587/1729-4061.2025.327907

SPEAKER RECOGNITION BY ULTRASHORT UTTERANCES

Bekbolat Medetov
PhD, Associate Professor*
Aigul Nurlankyzy
PhD Student
Department of Electronics, Telecommunications and Space Technologies**
Department of Space Engineering
Non-Profit Joint Stock Company "Almaty University of Power Engineering and Telecommunications named after Gumarbek Daukeyev"
Baytursynuli str., 126/1, Almaty, Republic of Kazakhstan, 050013
Timur Namazbayev
Master, Senior Lecturer
Department of Solid State Physics and Nonlinear Physics
Al-Farabi Kazakh National University
Al-Farabi ave., 71, Almaty, Republic of Kazakhstan, 050040
Ainur Akhmediyarova
PhD
Department of Software Engineering**
Kairatbek Zhetpisbayev
PhD
Department of Radio Engineering, Electronics and Telecommunications
Turan University
Satpayev str., 16A, Almaty, Republic of Kazakhstan, 050013
Ainur Zhetpisbayeva
Corresponding author
PhD, Associate Professor*
E-mail: aigulji@mail.ru
Aliya Kargulova
Senior Lecturer (Adviser)
Department of Electric Power Supply
S. Seifullin Kazakh Agrotechnical Research University
Zhenis ave., 62, Astana, Republic of Kazakhstan, 010011
*Department of Radio Engineering, Electronics and Telecommunications
L.N. Gumilyov Eurasian National University
Satbayev str., 2, Astana, Republic of Kazakhstan, 010008
**Satbayev University
Satbayev str., 22, Almaty, Republic of Kazakhstan, 050013

The object of this study is the accuracy of announcer identification based on short utterances.

To solve the task of speaker identification based on ultrashort speech utterances, a phoneme-by-phoneme approach to constructing voice models has been proposed within the framework of the study. The validity of this approach is based on the fact that short utterances usually contain a limited number of phonemes. In this regard, a hypothesis was put forward assuming that in order to increase the accuracy of announcer identification based on short utterances, it is necessary to analyze the sound of specific phonemes by different announcers.

The experiments involved speech recordings of monosyllabic words with corresponding phonemes, on the basis of which, using the ECAPA-TDNN neural network architecture, announcer voice models were constructed. The experimental studies showed that voice models constructed based on the sounds of only one model provide higher announcer identification accuracy compared to generalized models constructed based on all speech sounds.

It was also found that different phonemes provide different announcer identification accuracy. For example, with a speech signal duration of 2–3 seconds, the accuracy of announcer identification by the generalized model was 75 %. And the accuracy of announcer identification using a model built on the basis of only one phoneme "E", with the same input data, was 85 %, which is 10 percentage points higher than that of the generalized model

Keywords: announcer recognition, ultra-short utterances, phoneme-by-phoneme recognition, ECAPA-TDNN, phonemes of the Kazakh language

Received 04.02.2025
Received in revised form 25.03.2025
Accepted date 17.04.2025
Published date 29.04.2025

How to Cite: *Medetov, B., Nurlankyzy, A., Namazbayev, T., Akhmediyarova, A., Zhetpisbayev, K., Zhetpisbayeva, A., Kargulova, A. (2025). Speaker recognition by ultrashort utterances. Eastern-European Journal of Enterprise Technologies, 2 (9 (134)), 62–69. https://doi.org/10.15587/1729-4061.2025.327907*

1. Introduction

The development of digital technologies and the desire to design intelligent user interfaces put ever higher demands for personalized identification systems. In this context, voice biometrics is becoming one of the most promising areas, as it provides contactless, natural, and convenient interaction between a person and devices. Given the rapid spread of mobile platforms, IoT devices, and distributed computing systems, there is a need to devise solutions that would accurately identify a user with a minimum amount of available speech information.

Research in the field of voice biometrics has been conducted for quite a long time and methods of personal identification are constantly being improved. However, all known methods of announcer recognition show acceptable results only for long speech signals. In this regard, there is a need

to search for and devise new methods and approaches to announcer identification with short phrases, for example, when the announcer utters only a few words.

2. Literature review and problem statement

In [1], the performance of an announcer recognition system was improved by using the GMM-UBM neural network. Compared to similar tasks, this optimized classifier has fewer parameters, is trained faster, and has a higher accuracy level. For the voxceleb1 dataset with 1251 speakers, an accuracy of 94.33 % was achieved. However, this system is focused on long speech segments and is not designed to work with ultra-short utterances.

In [2], a study of emotional announcer recognition was conducted. In this paper, an attempt was made to extract emotional information in speech signals by simulating a mixed dataset of emotional speech for speaker recognition. Three datasets are considered, Language Emotional Speech Corpus, Crowdsourced Emotional Multimodal Corpus, Kharagpur Simulated Emotion Hindi Speech Corpus. The experimental results show that the performance of the system using mixed emotional speech improves speaker recognition. However, this model is not focused on the analysis of short speech fragments and does not take into account the phonemic structure of utterances, which limits its applicability under conditions of limited speech material.

In [3], a study was conducted on speaker recognition under noise conditions. The TIMIT dataset was used for the experiments, and the identification accuracy of 92.7 % was achieved. The proposed feature optimization method, applied to feature vectors with reduced dimensionality, led to a significant improvement in announcer recognition performance using various classification methods. Despite the high performance under noise conditions, the work did not consider solving the problem of announcer identification with ultra-short signal duration.

In [4], a new speaker identification method based on deep learning methods SVM and CNN was proposed. To conduct the speaker identification experiments, the LibriSpeech dataset with a gender split of 50 to 50 men and women was used, a total of 2484 audios were used, which is about 1000 hours of audio recordings. The recognition accuracy of 97.69 % was achieved. However, this method is focused on long speech segments and is also not designed to work with ultra-short utterances.

In [5], a feature extraction method SWFCC, stationary wavelet filter cepstral coefficients, is proposed in the announcer recognition task. The performance of the proposed SWFCC approach is evaluated on the TIMIT dataset. Experimental results using the Gaussian mixture model-universal background model (GMM-UBM) as a classifier show that SWFCC outperforms various feature extraction methods such as MFCC, PNCC, and GFCC. The results demonstrate the robustness of the model to noise but its performance in analyzing short utterances was not the subject of the study.

In [6], a study was conducted on speaker identification using the self-learning PNR-TDNN architecture, which is robust to strong noise and reverberation. The model was tested on the VoxCeleb1 dataset and showed improvements in the EER and MinDCF metrics compared to its peers. Despite the high accuracy under difficult acoustic conditions, the study does not evaluate the model's performance on short speech segments.

In [7], the CRET-1 and CRET-2 models for speaker recognition are proposed, based on the ECAPA-TDNN architecture and combining elements of CONV2D and ResNet. The models were trained on the large VoxCeleb2 corpus and demonstrated high announcer recognition accuracy (0.97828) and low error rate (EER=0.03612) when tested on the VoxCeleb1-O subset. However, the study is limited to segments of standard duration and does not evaluate the performance of the models on short or ultra-short speech fragments, which reduces their applicability under conditions of limited speech material.

In [8], a lightweight announcer recognition system based on timbre analysis is proposed. Timbre refers to the main properties of sound that allow listeners to distinguish between them. Using timbre properties such as brightness and sharpness, the system achieved a maximum accuracy of 78 % at the stage of announcer identification based on short utterances, the duration of which was about 2–3 seconds of pure speech. At the verification stage, the model maintains an accuracy of 76 % with an equal error rate (ERR) of 0.24. The speech data of 40 announcers from the LibriSpeech corpus is used to train the model. As we can see, the authors of the work failed to achieve high accuracy of announcer identification.

In [9], the study considers announcer recognition based on short utterances (1–2 seconds) using meta-learning and the mechanism of prototype networks. The authors note that conventional models show low accuracy under such conditions and propose an approach that improves the results on short utterances. The accuracy of announcer recognition for utterances 2 seconds long was about 74 %. The experiments were conducted using the VoxCeleb1 and VoxCeleb2 datasets, consisting of 1251 and 5994 announcers, respectively. Despite the use of promising methods based on meta-learning and the mechanism of prototype networks, the work also failed to achieve high accuracy of announcer recognition from short utterances.

Papers [10–12] consider the problem of announcer verification based on short utterances using the ECAPA-TDNN model and meta-learning. The authors note that existing methods require long audio fragments to achieve high announcer recognition accuracy and propose an approach that makes it possible to work with utterances from 2 to 5 seconds long. With an utterance length of 2 seconds, the verification accuracy was about 76 %. In the paper, the VoxCeleb1 dataset was used for experiments, which contains 37,720 complete utterances of 40 announcers. The ECAPA-TDNN neural network model shows very decent results in announcer identification based on long speech samples. But as the paper shows, that model also failed to provide high announcer identification accuracy based on short speech fragments.

Our review of the literature demonstrates that researchers mainly focus on general approaches to devising effective announcer recognition methods. All these approaches require long speech signals from announcers for their high-precision identification. And in those works where the problem of announcer identification by short fragments of speech was specifically studied, it was not possible to achieve high accuracy (it does not exceed 78 %).

Thus, the task of announcer identification by short phrases remains an unsolved problem today. The main issue is to increase the accuracy, reaching at least 90 %, in announcer recognition by short speech fragments. If we generalize the previous studies analyzed in our work, we can conclude that an attempt to solve the problem of announcer identification by short utterances is built around the combined use of speech features, the application of new learning strategies, the development of new neural network structures, etc.

However, to date, all these attempts have not yet yielded tangible results. In our opinion, when solving this problem, it is necessary to take into account that short speech utterances cannot contain all the sounds of the language (phonemes), accordingly, it is not possible to build one generalized human voice model using all possible phonemes, as is implemented in most existing methods of announcer identification.

If we proceed from the fact that the analyzed short speech signal will contain only a few phonemes, then it is assumed that for effective recognition of announcers it is also necessary to have a human voice model for each individual phoneme. It should be noted that different languages have different numbers of phonemes, for example, in British English there are 44 phonemes, in Kazakh there are 37 phonemes, and in Russian there are 42 phonemes. Thus, to implement this approach, one would need as many human voice models as there are phonemes, i.e., about 40–50 models. This is a lot; constructing so many different voice models for one person in practice is a difficult task.

But, on the other hand, there are several special phonemes that can be called universal, i.e., they are present in virtually all languages, these are the phonemes "A", "O", "E". Also, these phonemes are more often than other sounds in the word formation of human speech. In this regard, to build a human voice model, it is enough to limit ourselves to only these three phonemes.

In summary, speaker identification in short utterances should be performed using three voice models together. In other words, announcer identification is performed by analyzing how a person pronounces specific phonemes "A", "O", "E" in their utterances. However, in this case, another additional question arises. Do all three models have equal weight in making a decision? For example, it may happen that the models based on the phonemes "A" and "O" identify this person, while the model based on the phoneme "E" gives a negative decision; then what should be the final decision?

3. The aim and objectives of the study

The purpose of our study is to improve the accuracy of announcer identification based on short utterances. This will make it possible to determine which phonemes provide the highest recognition accuracy, as well as to select the optimal set of voice models with a limited speech signal length.

To achieve this goal, the following tasks are set:

– to determine the dependence of announcer recognition accuracy using a phoneme model on the duration of a speech signal;

– to conduct a comparative analysis of the identification ability of phonemes in announcer recognition.

4. The study materials and methods

4. 1. The object and hypothesis of the study

The object of our study is the accuracy of announcer identification based on short speech phrases. This paper puts forward a hypothesis that voice models built on individual phonemes have a higher accuracy in recognizing announcers in short phrases compared to generalized voice models that are built on all speech sounds.

It is assumed that voice models are neural networks trained on samples of voice data of various announcers. Our work does not aim to determine the correct structure of a neural network suitable for high-precision announcer identification. The principal task is to compare the accuracy of speaker recognition in short phrases using various voice models.

An analysis of the effectiveness of the phoneme-based approach to announcer recognition based on ultra-short speech utterances was carried out using the Kazakh language as an example, with an emphasis on determining the identification potential of individual vowel phonemes ("A", "O", "E").

The SpeechBrain library [13, 14], which is an open solution for speech processing tasks, was chosen as the basic platform for designing and conducting experiments to evaluate the effectiveness of phoneme-based announcer recognition. The pre-trained version of the ECAPA-TDNN model was used using standard tools and functions of the SpeechBrain library, which made it possible to effectively use the functions to implement announcer verification.

To conduct the experiments, test data were generated, which were recordings of the speech of 20 speakers. In this case, each announcer voiced 50 words containing each phoneme separately. Each recording included words in the Kazakh language with the structure: consonant-vowel-consonant (for example, "KET", "MEN", etc.).

The obtained audio data from each announcer was divided into the following two groups:

– voice samples used to construct a reference speaker model (for training the neural network);

– test data for assessing the accuracy of the trained neural network, not used to build the reference speaker model.

4. 2. Experimental procedure

The experiment was conducted with two different data sets. The first set included speech data containing only phonemes of one type. In the first stage of the study, words containing only one vowel phoneme (e.g., "A," "E," or "O") were analyzed and tested separately for each group of phonemes. The generated voice template also contained a single vowel phoneme corresponding to the analyzed group. During the experiment, as shown in Fig. 1, the amount of tested data varied from 3 to 24 words with a step of 3 for each set separately.

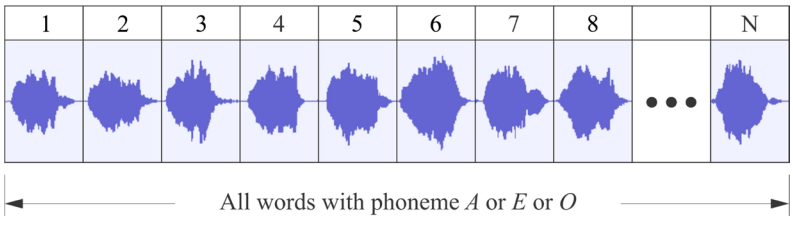

Fig. 1. Structure of test data for the first experiment

Fig. 1 illustrates the process of forming a sound stream containing only one phoneme. Each of the cells (1, 2, 3, and so on) contains different words, such as "Ket", "Men", "Kesh", with the total number of words being N=50. All selected words include two voiceless sounds and one voiced sound, which makes it possible to analyze the characteristics of the voice signal under conditions of minimal phonetic context.

The second data set was used for testing based on the combined speech fragments. At the second stage of the experiment, as shown in Fig. 2, words containing all three vowel phonemes

combined into a single sequence were used for testing. The length of the test sequences corresponded to the length used in the first experiment. The formed voice template included words containing all vowel phonemes (“A”, “O”, “E”).

During the experiment, the volume of test data varied from 3 to 24 words with a step of 3, with each test sample including words with three vowel phonemes in equal proportions. As can be seen from Fig. 2, the data are organized in such a way that words containing each of the phonemes alternate sequentially.

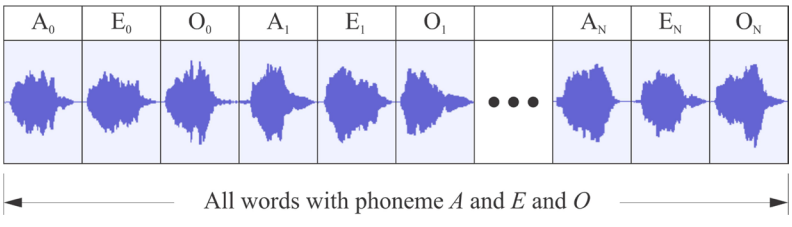

Fig. 2. Structure of test data for the second experiment

5. Results of analyzing the efficiency of phoneme-based announcer recognition

5. 1. Determining the dependence of announcer recognition accuracy using a phoneme model on the duration of a speech signal

To determine the efficiency of the phoneme-by-phoneme based approach, it is necessary to analyze the dependence of announcer recognition accuracy on the duration of a speech signal in the context of different phonemes. The end result of solving this problem is the type of functions that most accurately describe the dependence of announcer recognition errors by different models on the number of words in utterances (speech signal duration). The criterion for the accuracy of the solution to the problem in this case is the achieved accuracy of approximation of the experimental data, on the basis of which the functional dependence of announcer recognition error on the number of words in a speech signal is determined.

Our analysis of experimental data reveals that the dependence of the announcer recognition error on the number of words (duration of the speech signal) is nonlinear and is represented by a decreasing function. In this regard, four types of decreasing nonlinear functions were considered for the mathematical description of this pattern. These functions were used to approximate the empirical data for the most accurate mathematical representation of the identified dependence.

To analyze the data, we carried out approximation to identify patterns describing the obtained results. Four types of approximation functions were used in the study:

$$y = C \cdot x^n, \quad (1)$$

$$y = C \cdot n^x, \quad (2)$$

$$y = C \cdot e^{nx}, \quad (3)$$

$$y = a + b \cdot \log(x). \quad (4)$$

Based on the experimental data, the approximation accuracy was tested for each of the functions considered. The results of the estimate are given in Table 1, which shows the error values calculated for each of the approximation models.

Table 1

Value of the error in determining the significant parameters of approximating functions

Dataset type	$C \cdot x^n$	$C \cdot n^x$	$C \cdot e^{n \cdot x}$	$a+b \cdot \log(x)$
AEO	10.88 %	16.36 %	16.36	8.86 %
A	7.40 %	30.17 %	30.17 %	16.24 %
E	9.83 %	16.43 %	16.43 %	8.09 %
O	8.13 %	24.77 %	24.77 %	13.13 %

Analysis of the data from Table 1 reveals that functions (1) and (4) have the smallest error values, which indicates a high degree of their correspondence to the experimental data. However, to ensure uniformity of the mathematical description of all the processes under consideration, it is preferable to use one function. Function (1) was chosen as the main approximation model since it demonstrates a more uniform distribution of the error relative to all experimental results. The type of approximation function for the dependence of announcer recognition accuracy by voice for different types of phonemes and the generalized model is given in Table 2.

Table 2

The type of approximating function for the dependence of accuracy of announcer recognition by voice for different types of phonemes

No.	Words with phonemes	Approximating function
1	AOE	$y=0.762 \cdot x^{-0.350}$
2	A	$y=0.674 \cdot x^{-0.436}$
3	E	$y=0.762 \cdot x^{-0.504}$
4	O	$y=0.794 \cdot x^{-0.415}$

Fig. 3–6 show plots of approximation functions (1) for the phonemes “AEO”, “A”, “E”, “O”, constructed according to the data in Table 2.

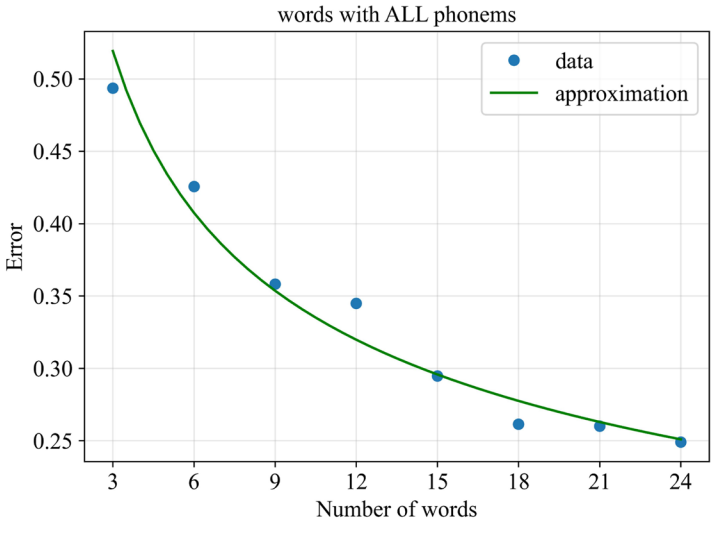

Fig. 3. Dependence of speaker recognition error when using a model constructed using all phonemes “A”, “E”, “O” on the number of words in the speech signal

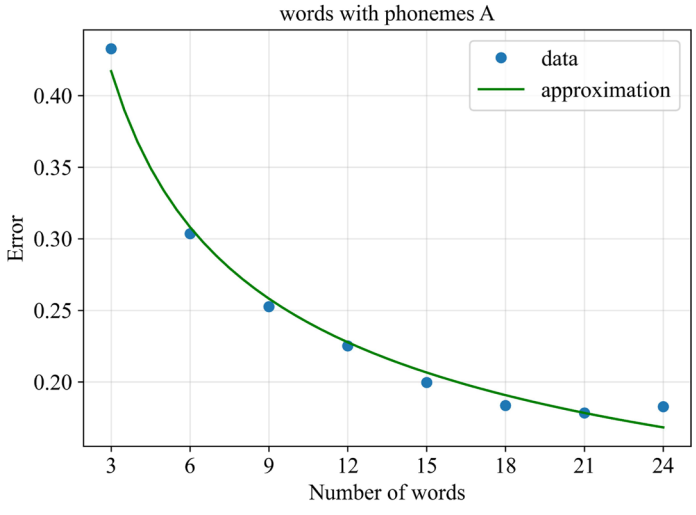

Fig. 4. Dependence of speaker recognition error when using a model constructed using one phoneme "A" on the number of words in the speech signal

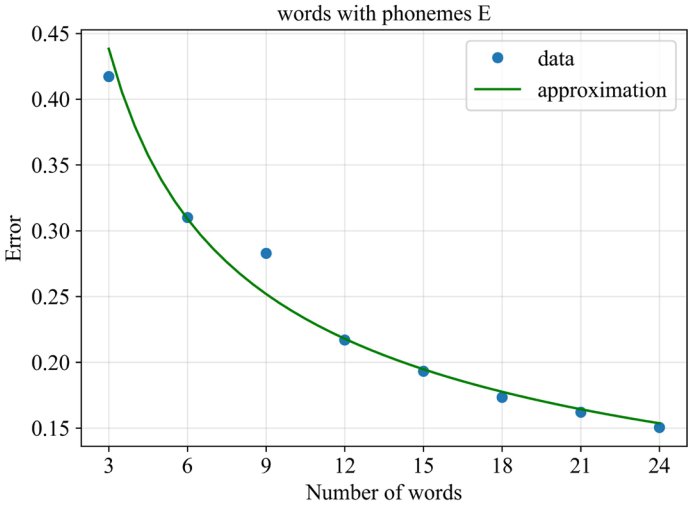

Fig. 5. Dependence of speaker recognition error using a model constructed using one phoneme "E" on the number of words in the speech signal

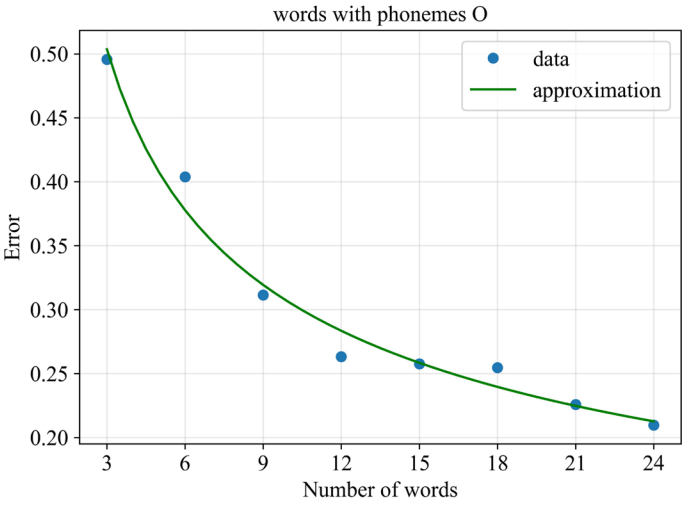

Fig. 6. Dependence of speaker recognition error using a model constructed using one phoneme "O" on the number of words in the speech signal

In Fig. 3–6, the experimental data are indicated by dots, and the green line shows the region of the approximation function corresponding to the range of experimental data (from 3 to 24 words). And in Table 2, as an example, all the approximation functions found for all three types of phonemes and the generalized model on utterances in the Kazakh language are given. In this case, the approximation error of the experimental data for different models took values from 7 % to 11 %. This is a fairly small approximation error, which makes it possible to state that the dependence of announcer recognition errors on the number of words in the speech signal is very well described by the functions given in Table 2.

5. 2. Conducting a comparative analysis of the identification ability of phonemes in announcer recognition

A comparative analysis of announcer recognition errors by different models was conducted in order to identify the model that yields the smallest error. The analysis was conducted based on the plots in Fig. 3–6, which show the dependences of speaker recognition errors on the number of words in the speech signal when using different voice models constructed using individual phonemes "A", "E", or "O", and a generalized model using all phonemes together. In order to compare announcer recognition errors by these different models, all of the above plots were represented on one diagram shown in Fig. 7.

The criterion for assessing the identification ability of different voice models is the recognition error value corresponding to a specific number of words in the speech signal. For example, for comparison, we can consider the recognition error when there are 24 words (the approximate duration of pure speech is about 2–3 seconds) in the speech signal. Then, from Fig. 7, we can see that the recognition error of the model for the phoneme "E" is 0.15, and for the generalized model, the error is 0.25. In other words, the recognition accuracy of the generalized model was 75 %, and the accuracy of the model for the phoneme "E" was 85 %, which is 7 percentage points higher than the maximum accuracy achieved in [8] for the same duration of the speech signal.

If we take into account the fact that all four voice models were built using the same ECAPA-TDNN neural network model with the same parameters, then we can say that the difference in speaker recognition errors is caused only by the difference in the training data set. In this case, Fig. 7 provides an answer to the question of which of the phonemes "A", "O", "E" or the generalized model "AOE" gives the best results in announcer recognition from short phrases. As can be seen from Fig. 7, all models built on individual phonemes provide a lower value of announcer recognition error compared to the generalized model. In this case, the best results are achieved when using a voice model formed on the basis of the phoneme "E".

It is obvious that when comparing the voice models of different announcers, their similarity should be minimal. Accordingly, the degree of similarity of the voice models of different announcers was also checked. The results of this analysis are shown in Fig. 8, which demonstrates average values of the percentage of similarity between different announcers.

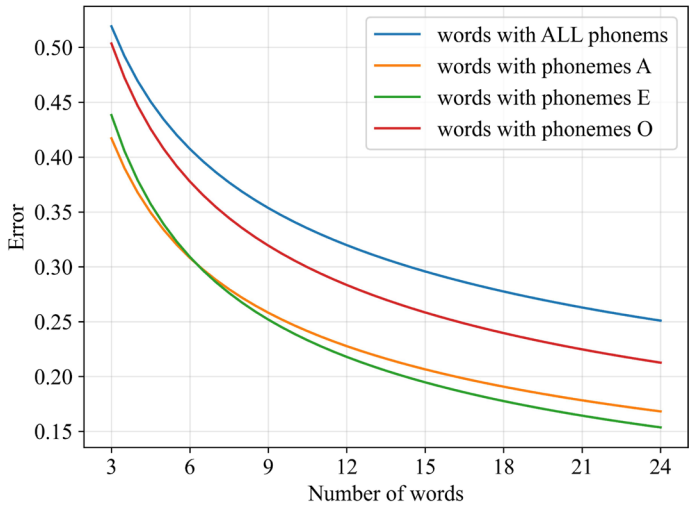

Fig. 7. Plots of speaker recognition errors depending on the number of words in the speech signal using different voice models

As can be seen from Fig. 8, both the phoneme-based and generalized models demonstrate that the percentage of similarity between different announcers does not exceed 23 %. This confirms that the proposed models effectively distinguish between the voices of different people. In addition, the figure shows that neither of the models – based on individual phonemes ("A", "O", "E") or the generalized model ("AOE") – has a clear advantage over the others in the task of distinguishing announcers.

Thus, analysis of the data in Fig. 8 reveals that the percentage of similarity of the voice models of different announcers does not depend on the choice of the phoneme-based or generalized model, which indicates the reliability of both methods in the task of announcer identification. At the same time, to determine the similarity of the voice characteristics of the same announcer, it is preferable to use a phoneme-by-phoneme based model since it provides higher recognition accuracy compared to the generalized model.

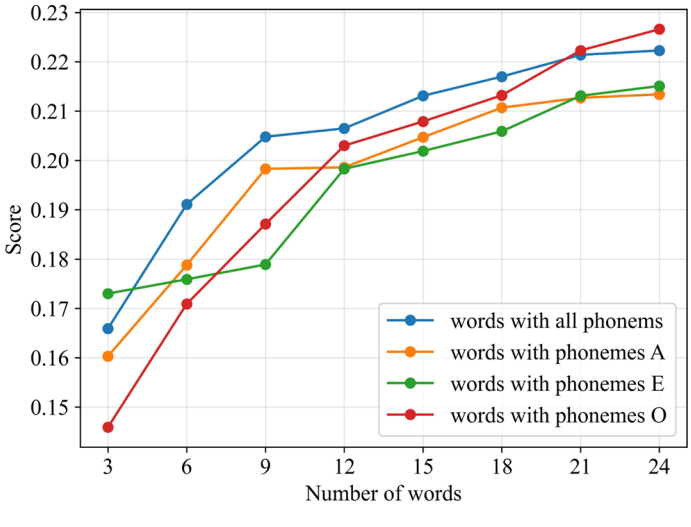

Fig. 8. Results of phoneme testing on Kazakh language data

Thus, the results of our analysis show that phoneme-by-phoneme based recognition does not provide significant differences in identifying different announcers. However, it is more effective when comparing voice samples of the same announcer. This indicates that phoneme-by-phoneme based recognition is appropriate for speaker verification where it is necessary to establish a correspondence between the presented voice sample and the standard.

6. Discussion of results based on investigating the effectiveness of phoneme recognition of announcers in short utterances

Our work has considered four different voice models built using the same neural network structure ECAPA-TDNN. Three of them were trained strictly on the basis of speech images of the corresponding one specific phoneme, and the fourth model, which is a generalized model, was trained on hybrid data including all phonemes.

First, as the results in Tables 1, 2, as well as the plots shown in Fig. 3–6, demonstrate, the dependence of the announcer recognition error on the number of words in the speech signal for all four models is very well described by a power function with a negative exponent. The decrease in the announcer recognition error with an increase in the signal duration (the number of words in speech) can be explained by the fact that with an increase in the number of words in the signal, the phonetic diversity of the training data increases. Obviously, this in turn leads to an increase in the ability of neural networks to generalize data. Knowing the law of change in announcer recognition accuracy from the number of words in the speech signal for each voice model makes it possible to predict their response at any speech duration. Consequently, it becomes possible to compare the effectiveness of the voice models under consideration with each other in a certain range of speech signal durations.

Secondly, according to the four plots shown in Fig. 7, we see that the generalized model has a higher announcer recognition error for any number of words in the speech signal compared to the models based on one specific phoneme. Thus, it has been established that for short utterances consisting of several words, the neural network model built on only one phoneme has a higher announcer identification accuracy compared to the model built without phoneme separation. Accordingly, it can be argued that the original hypothesis that voice models built on individual phonemes have a higher accuracy in announcer recognition in short phrases compared to generalized voice models has its experimental confirmation. The higher accuracy of announcer recognition by voice models built on specific phonemes compared to generalized models can be explained by several factors. Firstly, the generalized model must be trained on all possible speech sounds, and accordingly, the volume of the training sample must be large enough. Secondly, a phonemic voice model requires only one phoneme to be trained and therefore requires much less training data. Thirdly, short utterances usually have a limited number of phonemes, so phonemic models can more accurately capture the features of speech sounds.

From the same plots shown in Fig. 7, it can be seen that models based on specific phonemes have different announcer recognition accuracy. If we focus only on the results of this work, where only speech data in the Kazakh language were used in the experiment, it is clear that the phoneme "E"

has the best identification ability. In this work, the reason why this phoneme has the highest accuracy in announcer identification was not analyzed. Nevertheless, knowledge of the difference in announcer identification accuracy by different phoneme models could make it possible to take the right decision in cases where announcer recognition is carried out using several such models.

The advantage of the proposed approach is the ability to build a variety of specialized voice models based on individual phonemes. Unlike universal methods [1–7], which use generalized voice models, our approach makes it possible to more accurately adapt to the features of sounds in a particular language, specifically, Kazakh. This became possible owing to a targeted analysis of the pattern of changes in the accuracy of announcer identification depending on the duration of speech signals using voice models constructed on the basis of individual phonemes and limiting the phonetic composition, which was not considered in the aforementioned works.

At the same time, the study has a number of limitations. Firstly, it analyzes only three vowel phonemes. This provides a partial picture and requires expanding the study to the entire phonemic composition of the Kazakh language. Secondly, only one architecture was used – ECAPA-TDNN. Perhaps, when using other models, the results could be different.

Among the disadvantages, one can note the lack of automatic phoneme recognition in continuous speech, which limits the practical implementation of the proposed method in real systems. It is also necessary to take into account that the construction of separate models for each phoneme requires additional computing resources. Thus, phoneme-by-phoneme announcer recognition based on ultra-short utterances demonstrates stability, practical applicability, and the potential for expansion, especially in the context of language specificity and limited data.

The fact that the phoneme model is capable of providing higher accuracy of announcer recognition compared to the generalized model makes this method promising. In the future, additional research is required to answer a number of important questions. For example, is it possible to implement language-independent announcer recognition using this approach? It is also necessary to analyze what final accuracy such a system can provide, taking into account the need to recognize individual phonemes in continuous speech.

7. Conclusions

1. It has been found that the announcer recognition accuracy depending on the speech signal duration (number of words) for all four voice models is very well described by a decreasing power function. The relative approximation error of the power function in our case lies in the range from 7 % to 11 %. As is known, the value of the average approximation error of up to 15 % indicates a well-chosen model describing the pattern of experimental data.

2. A comparative analysis of the announcer identification error by different voice models has revealed that models based on individual phonemes have higher accuracy compared to the generalized model. For example, for an utterance consisting of 24 words (corresponding to a signal duration of 2–3 seconds of pure speech), the recognition error of the generalized model turned out to be approximately 0.25, and for phoneme models “A”, “O”, and “E”, the error values are 0.18, 0.21, and 0.15, respectively. These numbers show that the voice model built on the phoneme “E” has the highest announcer identification accuracy, significantly exceeding the generalized model. At the same time, the model built on the phoneme “E” for 2–3 seconds of shear speech provided announcer recognition accuracy of 85 %, which is 7 percentage points higher than the maximum accuracy achieved in previous studies.

Conflicts of interest

The authors declare that they have no conflicts of interest in relation to the current study, including financial, personal, authorship, or any other, that could affect the study, as well as the results reported in this paper.

Funding

This research is funded by the Science Committee of the Ministry of Science and Higher Education, the Republic of Kazakhstan (Grant No. AP19678995) “Development of a speaker recognition method using deep neural networks for ultrashort duration of pure speech”.

Data availability

The data will be provided upon reasonable request.

Use of artificial intelligence

The authors used artificial intelligence technologies within acceptable limits to provide their own verified data, which is described in the research methodology section.

References

1. Sharif-Noughabi, M., Razavi, S. M., Mohamadzadeh, S. (2025). Improving the Performance of Speaker Recognition System Using Optimized VGG Convolutional Neural Network and Data Augmentation. International Journal of Engineering, 38 (10), 2414–2425. https://doi.org/10.5829/ije.2025.38.10a.17
2. Tomar, S., Koolagudi, S. G. (2025). Blended-emotional speech for Speaker Recognition by using the fusion of Mel-CQT spectrograms feature extraction. Expert Systems with Applications, 276, 127184. https://doi.org/10.1016/j.eswa.2025.127184
3. Chauhan, N., Isshiki, T., Li, D. (2024). Enhancing Speaker Recognition Models with Noise-Resilient Feature Optimization Strategies. Acoustics, 6 (2), 439–469. https://doi.org/10.3390/acoustics6020024
4. Kohler, O., Imtiaz, M. (2025). Investigation of Text-Independent Speaker Verification by Support Vector Machine-Based Machine Learning Approaches. Electronics, 14 (5), 963. https://doi.org/10.3390/electronics14050963

5. Missaoui, I., Lachiri, Z. (2025). Stationary wavelet Filtering Cepstral coefficients (SWFCC) for robust speaker identification. Applied Acoustics, 231, 110435. https://doi.org/10.1016/j.apacoust.2024.110435
6. Zhang, X., Tang, J., Cao, H., Wang, C., Shen, C., Liu, J. (2025). A Self-Supervised Method for Speaker Recognition in Real Sound Fields with Low SNR and Strong Reverberation. Applied Sciences, 15 (6), 2924. https://doi.org/10.3390/app15062924
7. Li, P., Hoi, L. M., Wang, Y., Yang, X., Im, S. K. (2025). Enhancing Speaker Recognition with CRET Model: a fusion of CONV2D, RESNET and ECAPA-TDNN. EURASIP Journal on Audio, Speech, and Music Processing, 2025 (1). https://doi.org/10.1186/s13636-025-00396-4
8. Ohi, A. Q., Mridha, M. F., Hamid, M. A., Monowar, M. M., Lee, D., Kim, J. (2020). A Lightweight Speaker Recognition System Using Timbre Properties. arXiv. https://doi.org/10.48550/arXiv.2010.05502
9. Kye, S. M., Jung, Y., Lee, H. B., Hwang, S. J., Kim, H. (2020). Meta-Learning for Short Utterance Speaker Recognition with Imbalance Length Pairs. Interspeech 2020. https://doi.org/10.21437/interspeech.2020-1283
10. Wang, W., Zhao, H., Yang, Y., Chang, Y., You, H. (2023). Few-shot short utterance speaker verification using meta-learning. PeerJ Computer Science, 9, e1276. https://doi.org/10.7717/peerj-cs.1276
11. Chen, Y., Zheng, S., Wang, H., Cheng, L., Zhu, T., Huang, R. et al. (2025). 3D-Speaker-Toolkit: An Open-Source Toolkit for Multimodal Speaker Verification and Diarization. ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 1–5. https://doi.org/10.1109/icassp49660.2025.10888389
12. Desplanques, B., Thienpondt, J., Demuynck, K. (2020). ECAPA-TDNN: Emphasized Channel Attention, Propagation and Aggregation in TDNN Based Speaker Verification. Interspeech 2020. https://doi.org/10.21437/interspeech.2020-2650
13. Ravanelli, M., Parcollet, T., Moumen, A., de Langen, S., Subakan, C., Plantingaet, P. al. (2024). Open-Source Conversational AI with SpeechBrain 1.0. arXiv. https://arxiv.org/abs/2407.00463
14. Ravanelli, M., Parcollet, T., Plantinga, P., Rouhe, A., Cornell, S., Lugosch, L. et al. (2021). SpeechBrain: A General-Purpose Speech Toolkit. arXiv. https://arxiv.org/abs/2106.04624