

МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ
АКАДЕМИЯ ПРОБЛЕМ КАЧЕСТВА РОССИЙСКОЙ ФЕДЕРАЦИИ
РОССИЙСКАЯ АКАДЕМИЯ КОСМОНАВТИКИ им. К. Э. ЦИОЛКОВСКОГО
ФЕДЕРАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ ЦЕНТР «ИНФОРМАТИКА И УПРАВЛЕНИЕ» РАН
(ВЫЧИСЛИТЕЛЬНЫЙ ЦЕНТР РАН имени А. А. ДОРОДНИЦЫНА)
АО «УПКБ "ДЕТАЛЬ"», АО «РАДИОЗАВОД»
АО «ФНПЦ «ПО "СТАРТ" имени М. В. ПРОЦЕНКО», АО «НИИЭМП»
АО «ПНИЭИ», АО «НИИФИ», АО «НПП "РУБИН"»
ЕВРАЗИЙСКИЙ НАЦИОНАЛЬНЫЙ УНИВЕРСИТЕТ имени Л. Н. ГУМИЛЕВА
ПЕНЗЕНСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ



ТРУДЫ
МЕЖДУНАРОДНОГО СИМПОЗИУМА
НАДЕЖНОСТЬ И КАЧЕСТВО

Том 1

*Посвящается году единства народов России и 65-летию
со дня первого полета Юрия Гагарина в космос*

ПЕНЗА 2026

Труды Международного симпозиума НАДЕЖНОСТЬ И КАЧЕСТВО. – 2026. – Т. 1. – 467 с.

Представлены доклады XXXI Международного симпозиума «Надежность и качество», посвященного году единства народов России и 65-летию со дня первого полета Юрия Гагарина в космос, проходившего с 25 по 31 мая 2026 г. в г. Пензе. Рассмотрены актуальные проблемы теории и практики повышения надежности и качества; эффективности внедрения инновационных и информационных технологий в фундаментальные научные и прикладные исследования, образовательные и коммуникативные системы и среды, экономику и юриспруденцию; методов и средств анализа и прогнозирования показателей надежности и качества приборов, устройств и систем, а также анализа непараметрических моделей и оценки остаточного ресурса изделий двойного назначения; ресурсосбережения; проектирования интеллектуальных экспертных и диагностических систем, систем управления и связи, интерактивных, телекоммуникационных сетей и сервисных систем; экологического мониторинга и контроля состояния окружающей среды и биологических объектов; исследования физико-технологических процессов в науке, технике и технологиях для повышения качества выпускаемых изделий радиопромышленности, приборостроения, аэрокосмического и топливно-энергетического комплексов, электроники и вычислительной техники и др.

Издание предназначено для студентов, аспирантов, ученых и научных сотрудников, заинтересованных в исследовании надежности и качества эффективных комплексных систем.

УДК 621.396.6:621.315.616.97:658:562

Оргкомитет благодарит за поддержку в организации и проведении Международного симпозиума и издания настоящих трудов Министерство науки и высшего образования РФ, Правительство Пензенской области, Академию проблем качества РФ, Российскую академию космонавтики имени К. Э. Циолковского, Российскую инженерную академию, ФИЦ ИУ РАН (Вычислительный центр РАН имени А. А. Дородницына), АО «УПКБ “ДЕТАЛЬ”», АО «НИИФИ», АО «ПНИЭИ», АО «НПП “РУБИН”», АО «РАДИОЗАВОД», АО «ППО “ЭЛЕКТРОПРИБОР”», АО «НИИЭМП», АО «ФНПЦ «ПО “СТАРТ” имени М. В. Проценко», АО «НИКИРЭТ», ООО «Экспериментальная мастерская “Наука-Софт”», Евразийский национальный университет имени Л. Н. Гумилева, Пензенский государственный университет.

Издание зарегистрировано в Российском индексе научного цитирования (РИНЦ) с 2005 г.

Редакционная коллегия:

Юрков Н. К. – главный редактор;
Трусов В. А. – ответственный секретарь;
Гуляков А. Д., Васин С. М., Сыдыков Е. Б., Абрамов О. В.,
Артамонов Д. В., Безродный Б. Ф., Бецков А. В., Дивеев А. И., Иофин А. А.,
Кашищанов В. А., Кемалов Б. К., Острейковский В. А., Пиганов М. Н.,
Роберт И. В., Романенко Ю. А., Савченко В. П., Садыхов Г. С.,
Халютин С. П., Шахнов В. А., Якимов А. Н.

УДК 004.93'11

Arfan¹ A., Burukina² I.P., Gorbunov³ N.A.

¹ Independent researcher, Morocco

^{2,3} Penza State University, Penza, Russia

¹ arfaneadam011@gmail.com

² burukinairina@gmail.com

³ nik.gorbunov.02022002@mail.ru

TARGET VOICE EXTRACTION FROM NOISY AUDIO SIGNALS: A COMPARISON OF NEURAL NETWORK AND ENERGY-BASED APPROACHES

Abstract. This paper addresses the relevant problem of extracting target speech from noisy audio signals. We analyze two approaches to solving this problem: a method based on the deep neural network ECAPA-TDNN using speaker embeddings for speaker identification, and a classical energy-based method relying on binary masking in the time-frequency domain. The architectural features of the neural network approach, including the weighted time averaging mechanism, are described, and a mathematical justification for both methods is provided. A comparison of the characteristics of the considered algorithms is performed, demonstrating the advantages of neural network methods in extraction accuracy at the cost of higher computational complexity.

Keywords: neural network, recognition, target speech, ECAPA-TDNN, weighted time averaging, noise reduction

Арфан¹ А., Бурукина² И.П., Горбунов³ Н.А.

¹ Независимый исследователь, Марокко

^{2,3} Пензенский государственный университет, Пенза, Россия

¹ arfaneadam011@gmail.com

² burukinairina@gmail.com

³ nik.gorbunov.02022002@mail.ru

ВЫДЕЛЕНИЕ ЦЕЛЕВОГО ГОЛОСА НА ФОНЕ ЗАШУМЛЕННОГО АУДИОСИГНАЛА: СРАВНЕНИЕ НЕЙРОСЕТЕВОГО И ЭНЕРГЕТИЧЕСКОГО ПОДХОДОВ

Аннотация. Рассматривается актуальная задача выделения целевого голоса на фоне зашумленного аудиосигнала. Проведен анализ двух подходов к решению этой проблемы: метода на основе глубокой нейросети ECAPA-TDNN, использующего спикерные эмбединги для идентификации диктора, и классического энергетического метода, основанного на бинарной маскировке в частотно-временной области. Описаны архитектурные особенности нейросетевого подхода, включая механизм взвешенного усреднения по времени, а также представлено математическое обоснование обоих методов. Выполнено сравнение характеристик рассмотренных алгоритмов, демонстрирующее преимущества нейросетевых методов в точности выделения при более высокой вычислительной сложности.

Ключевые слова: нейронная сеть, распознавание, целевой голос, ECAPA-TDNN, взвешенное усреднение по времени, шумоподавление

Introduction. The problem of extracting target speech from background noise and other speech sources is one of the fundamental tasks in modern audio signal processing systems. Despite the popularity of voice interfaces, automatic translation systems, and intelligent assistants, a key issue that severely limits the effectiveness of such systems remains the efficient extraction of target speech. In real acoustic scenarios – whether a busy street, office space, car, or echoic room – the audio signal is rarely isolated. Background noise, other people's voices, and reverberations are added to the speaker's speech, hindering correct perception and processing by algorithms.

Target speech extraction is the process of processing an audio signal to isolate the speech of a desired speaker from the overall recording while reducing the influence of noise and extraneous sounds. This task is important for improving the accuracy of speech recognition systems and enhancing the overall performance of voice technologies.

Currently, digital signal processing methods, statistical models, and neural network approaches are used to solve this problem. Deep neural networks, such as convolutional and recurrent models as well as attention mechanisms, have proven particularly effective. They allow a significant increase in speech extraction quality. However, problems remain regarding model robustness to new types of noise and their ability to operate under various conditions.

Target speech extraction algorithm based on ECAPA-TDNN

One promising approach to target speech extraction is the use of specific voice characteristics of a particular person. These characteristics allow the system to distinguish the desired speaker's voice from other voices and noise. For obtaining these characteristics, the ECAPA-TDNN neural network model is used [1-3]. A short reference recording is used, from which the system identifies the speaker and extracts their speech from the audio stream.

The ECAPA-TDNN architecture, implemented for example in the SpeechBrain library, is trained on the VoxCeleb dataset [4] and used for speaker

identification. The model processes a fixed-length audio signal and transforms it into an embedding – a numerical representation of the voice as a feature vector. The embedding reflects individual voice characteristics and is used for comparing speakers.

The algorithm's operation can be organized as follows:

Obtaining a reference voice representation. A vector representation of the voice is computed from the target speaker's recording, used as a template.

Splitting the audio signal into segments. The original recording is divided into fixed-length temporal segments.

Analysis of voice characteristics of segments. For each segment, a vector representation is computed using the ECAPA-TDNN model.

Comparison with the reference voice. Segment embeddings are compared with the reference vector using vector using cosine similarity.

Selection of relevant segments. Segments exceeding a given similarity threshold are combined into the output audio signal

Formation of the output audio signal. The selected segments are concatenated into the final signal.

The ECAPA-TDNN model transforms the input audio signal into a compact voice vector representation through stepwise processing and aggregation of temporal features. A key element of the architecture is the weighted time averaging mechanism, which allows determining which speech segments are most informative for speaker identification [5].

Let h_t be the feature vector obtained at the last convolutional layer of the model for time step t . For each step and time channel, an importance value is computed according to the following formula:

$$e_{t,c} = v_c^T f(W h_t + b) + k_c$$

Where W , b and other model parameters determine the contribution of each time step and channel to the final representation.

The obtained values are normalized over time using the softmax function:

$$\alpha_{t,c} = \frac{\exp(e_{t,c})}{(\sum_{\tau=1}^T \exp(e_{\tau,c}))}$$

The coefficients $\alpha_{t,c}$ define the relative importance of each time segment in forming the final voice representation.

Based on these coefficients, weighted statistics – the mean and standard deviation of the features – are computed [6]:

$$\mu_c = \sum_{t=1}^T \alpha_{t,c} h_{t,c}$$

$$\tilde{\sigma}_c = \sqrt{\sum_{t=1}^T \alpha_{t,c} h_{t,c}^2 - \tilde{\mu}_c^2}$$

The vectors are concatenated to form a fixed-length final voice vector representation. The resulting vector reflects the individual voice characteristics.

To compare two voices, the cosine similarity between their corresponding vector representations is used:

$$\text{sim}(x, y) = \frac{(x \cdot y)}{(\|x\| \|y\|)}$$

Values close to 1 indicate a high degree of voice similarity, while lower values indicate differences between speakers.

The method provides high accuracy even under strong interference and robustness to overlapping voices, but requires significant computational resources and a trained model.

Energy masking in the STFT representation

One widely used approach for extracting target speech signals is the binary masking method in the time-frequency domain. The signal is transformed into a spectral representation [7] using the short-time Fourier transform, after which the energy is estimated for each time frame. Low-energy segments are suppressed because they are assumed to correspond mainly to noise or uninformative components. This approach belongs to single-channel heuristic methods and can be used as a simple and computationally inexpensive way for preliminary speech signal extraction [8].

The signal is transformed into the time-frequency domain using STFT. A matrix of complex coefficients is formed, where the magnitude is used for analysis and the phase is preserved for signal reconstruction.

For each time frame, the average spectral energy is calculated as the mean of the spectral magnitude over all frequency components. This average spectral energy is used to distinguish speech from noise segments. Thus, the energy characteristic serves as a feature for extracting speech segments.

Based on the energy distribution, a threshold (e.g. the 75th percentile) is selected. Frames above the threshold are considered speech frames and form a binary mask that is applied to all frequencies. The mask is applied to the spectrum, after which the inverse STFT is performed to reconstruct the time-domain signal. This method is simple and computationally inexpensive but is sensitive to the choice of threshold and does not account for speaker individuality, extracting only high-energy segments.

Results. For objective comparison of the effectiveness of the mentioned methods, a series of experiments was conducted under various conditions, allowing quantitative assessment of speech extraction quality. For this purpose, a synthetically generated database was used, including clean speech signals from the LibriSpeech and VoxCeleb datasets, as well as background noises of various types. Each test signal was formed as the sum of the target signal and several noise sources with controlled gains and a signal-to-noise ratio ranging from -10 to +10dB. Experiments were performed while varying the number of noise sources.

The algorithms were implemented in Python using the Librosa library for the energy-based approach and the pre-trained ECAPA-TDNN model from the SpeechBrain repository for the neural network approach. To evaluate extraction quality, the following metrics were used: scale-invariant signal-to-noise ratio (SI-SNR), short-time objective intelligibility (STOI), and probability of correct consonants (PCC). Experiments were conducted on a computing server with GRU, and the statistical significance of differences was assessed using Student's t-test. The experimental results are presented in Table.

The results of the comparative analysis

Table

Condition	Method	SI-SNR	STOI	PCC
Clean signal	Energy-based	18,2	0,94	-
	ECAPA-TDNN	19,8	0,97	0,99
Noise (SNR=0 dB)	Energy-based	6,4	0,62	-
	ECAPA-TDNN	9,8	0,78	0,92
Speech + speech	Energy-based	4,1	0,48	-
	ECAPA-TDNN	7,3	0,65	0,85

Analysis of the obtained results showed a statistically significant superiority of the ECAPA-TDNN neural network method across all metrics, especially in challenging conditions with low SNR and the presence of interfering speech sources.

This paper has considered methods for target speech extraction based on speaker embeddings and energy analysis. Such methods provide more accurate speaker extraction but require complex models, whereas the energy approach is simpler to implement but inferior in accuracy. The obtained results can serve as a basis for further development of target speech extraction algorithms.

REFERENCES

1. Liu S., Song Z., He L. Improving ecapa-tdnn performance with coordinate attention//Journal of Shanghai Jiaotong University (Science). 2024. P. 1-7.
2. Wang D. et al. ECAPA-TDNN Based Depression Detection from Clinical Speech//Interspeech. 2022. P. 3333-3337.
3. Zhang E., Wu Y., Tang Z. SC-EcapaTdn: ECAPA-TDNN with separable convolutional for speaker recognition//International Conference on Intelligence Science. Cham: Springer Nature Switzerland. 2024. P. 286-297.
4. Xiao X. et al. Microsoft speaker diarization system for the voxceleb speaker recognition challenge 2020//ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2021. P. 5824-5828.
5. Togneri R., Pullella D. An overview of speaker identification: Accuracy and robustness issues//IEEE circuits and systems magazine. 2011. V.11(2). P. 23-61.
6. Wang Y. M., Luo Y. Integration of correlations with standard deviations for determining attribute weights in multiple attribute decision making//Mathematical and Computer Modelling. 2010. V.51(1-2). P. 1-12.
7. Tikhonov E., Sneps-Snepp M. Introduction to signal processing: spectral representation//International Journal of Open Information Technologies. 2019. V.7(4). P. 53-60.
8. Singh M. K. Feature extraction and classification efficiency analysis using machine learning approach for speech signal//Multimedia Tools and Applications. 2024. V.83(16). P. 47069-47084.