Кібернетика та комп'ютерні технології

УДК 81'1:003:004.932.2(045) DOI:10.34229/2707-451X.25.4.5

І.А. БЕЗВЕРБНИЙ, К.П. СОСНЕНКО

СЕМІОТИЧНИЙ ПІДХІД У СТВОРЕННІ ФОНЕМНОЇ МОДЕЛІ МОВНОГО СИГНАЛУ

Розглянуто семіотичний підхід до побудови фонемної моделі аналізу мовного сигналу. Запропоновано подання сигналу як послідовності знакових одиниць, утворених різницею миттєвих параметрів частоти та амплітуди. Такий формат дозволяє відобразити як фізичні, так і функціональні характеристики мовлення, що робить модель інтерпретованою та придатною для пояснюваного штучного інтелекту. Особливу увагу приділено нормалізації частот за допомогою хроматичного звукоряду та mel-шкали, що узгоджується з психоакустичними властивостями слуху. Запропонована модель може бути використана у дослідженнях фонології, комп'ютерної лінгвістики та системах штучного інтелекту.

Ключові слова: *мовний сигнал, лінгвістична структура, фонемна модель, інтерпретованість, семіотична репрезентація, рекурентна нейронна модель.*

Вступ. Незважаючи на значні успіхи, досягнуті в процесі розвитку сучасних систем мовного аналізу, залишається низка фундаментальних проблем, пов'язаних із нестабільністю акустичних ознак. Серед проблем – надзвичайна залежність від величезних обсягів даних, низька стійкість до акустичних шумів та варіацій мовлення у реальних умовах, складність узагальнення на непередбачені дані та ефективна обробка рідкісних слів. Крім того, відсутність інтерпретованості моделей ускладнює діагностику помилок, а значні обчислювальні витрати перешкоджають їхньому широкому розгортанню на пристроях з обмеженими ресурсами.

Ідея спрощення обчислень може полягати у створенні апріорних методів мовного аналізу. Скажімо, існуючі підходи розглядають мовний сигнал як сукупність ознак, отриманих із використанням спектрального аналізу, частотного перетворення, спектрограм або інших акустичних методів, які орієнтовані на статистичне узгодження акустичних шаблонів із лінгвістичними одиницями. Тим часом, практичні дослідження показали, що фонемна ідентифікація залишається стабільною навіть при абстрагуванні від точних фізичних параметрів мовного сигналу. Тобто апріорні значення фізичних параметрів, у нашому випадку частоти, так само як і амплітуди, доносять до слухового апарату людини однозначну ідентифікацію мовного сигналу. Як результат з'явилася ідея про представлення мовного сигналу, який функціонує не лише як акустичний потік, але і як послідовність структурованих звукових знаків певної частоти і амплітуди. Тому фонема може бути представлена як структура знакової системи, де звукові одиниці функціонують не як акустичні явища, а як функціонально організовані елементи мовної комунікації.

Виходячи з цього, наступним етапом формалізації мовного сигналу може бути створення алфавіту стандартизованих значень амплітуди і частоти, на основі якого з'явиться можливість оперувати структурами, які є узгодженими з фонемною системою мови з метою побудови інтерпретованих шаблонів, які відповідають фонемам як знаковим одиницям і забезпечують зв'язок між акустичною формою та фонологічною функцією.

Ступінь розробки. Проблематика автоматичного розпізнавання мовлення досліджується вже понад півстоліття, за цей час було створено низку моделей, які демонструють високу ефективність у задачах транскрипції мовлення у текст. Перші успішні спроби автоматичного розпізнавання мовлення були пов'язані з використанням моделей прихованих марковських процесів у поєднанні з гауссовими сумішевими моделями [1]. Ці системи здійснювали послідовне узгодження коротких акустичних фреймів із фонемними мітками, використовуючи оцінювання ймовірності та контекстні моделі. З переходом до глибокого навчання з'явились DNN-HMM гібриди, де ймовірності були апроксимовані глибокими нейромережами [2]. Пізніше виникли end-to-end моделі з використанням Connectionist Temporal Classification (CTC) [3], attention-based моделей, а також трансформери для мовлення [4].

У більшості сучасних систем АРМ фонема – це мітка у навчальній вибірці. Деякі роботи намагаються використати фонемні послідовності як проміжний рівень між сигналом і текстом (наприклад, phoneme-to-grapheme models (P2G) [5] або [6]), але у цих підходах відсутній аналіз внутрішньої форми фонеми як послідовності ознак або знаків нижчого порядку.

Окремі спроби ввести пояснюваність через attention-механізми, alignment та фонетичні embeddings (наприклад, wav2vec 2.0 [4] у поєднанні CTC alignment [3]), фокусуються на локалізації, а не на структурі фонемного знака. Іншими словами всі ці моделі працюють з випадковими акустичними ознаками (мел-частотна шкала, спектрограми, filter banks), і не використовують згадані вище апріорні моделі.

Концепція кореляції частот мовного сигналу з частотами дванадцятитонового хроматичного звукоряду [7–10] сформувалася на підставі аналізу анатомо-фізіологічної організації слухового апарату людини і є ідейно-теоретичним підґрунтям для розробки підходу до нормалізації мовного сигналу за частотними характеристиками цього звукоряду.

Ідея нормалізації мовного сигналу по частотам дванадцятитонового хроматичного звукоряду. Як зазначалося вище фонемна ідентифікація залишається стабільною при абстрагуванні від точних фізичних параметрів мовного сигналу. Іншими словами, нормалізація миттєвих частот до апріорного сталого ряду значень практично не впливає на ідентифікацію фонеми. Виходячи з уявлень про будову людського вуха, резонанс базилярної мембрани забезпечує вибірковість звуків в рамках звучання фонеми подібно до послідовності музичних звуків у музичній фразі [7, 8]. Тому пропонується нормалізація миттєвих частот мовного сигналу хроматичним звукорядом, обмеженим так званим мовним ядром, тобто діапазоном частот, які використовує людина у мовному сигналі (мовне ядро), а саме діапазоном 0,3–3,4 кГц. Миттєві частоти, які не потрапляють у зазначений діапазон, розглядаються як шуми і прибираються з сигналу або конвертуються до кінцевих значень діапазону.

У хроматичному звукоряді кожен півтон відповідає збільшенню частоти на множник $2^{1/12} \approx 1.05946$. Це дозволяє побудувати логарифмічну сітку нормалізації у межах частотного діапазону, притаманного мовленню (наприклад, 100 Гц – 8000 Гц). Кількість півтонів N, що вкладаються у цей діапазон, визначається за формулою:

$$N = 12 \cdot \log_2 \left(\frac{f_{\max}}{f_{\min}} \right). \quad (1)$$

Для діапазону 100–8000 Гц отримаємо приблизно 76 дискретних кроків. Кожному кроку відповідає окремий символ словника комбінацій стандартизованих значень частоти і амплітуди. Таким чином, зміна миттєвої частоти сигналу подається як дискретна послідовність, що має чітку інтерпретацію у музично-психоакустичних категоріях.

Амплітудна складова сигналу також підлягає нормалізації – зазвичай у менше число рівнів (наприклад, 20–30), з можливим використанням логарифмічного масштабування або нормалізації до уніфікованої шкали. Комбінація коду частоти (перша літера) та коду амплітуди (друга літера) формує дволітерний символ, що стає одиницею аналізу моделі.

Можна також розглянути альтернативний підхід. З огляду на те, що людське сприйняття частоти є нелінійним, можливо застосувати альтернативну нормалізацію миттєвої частоти у mel-шкалі, яка лінійна у низьких частотах і логарифмічна у високих. Перехід від частоти f у Гц до значення у шкалі mel здійснюється за формулою:

$$mel(f) = 2595 \cdot \log_{10}\left(1 + \frac{f}{700}\right). \quad (2)$$

Після переведення у шкалу mel частотний діапазон розбивається на рівномірні інтервали (наприклад, 64 або 80), які потім кодуються символами аналогічно до хроматичного підходу.

Такий метод узгоджується із принципами аудіоаналізу в сучасних АРМ-системах і може виявитися більш точним для інтерпретації сегментів мовного сигналу, що знаходяться у нижньому частотному регістрі.

Особливістю практичного використання зазначених підходів стало використання не миттєвих частот, а миттєвих зворотніх величин – довжин хвиль. Довжина хвилі може бути визначена через число вибірок, які обчислюються на ділянці з незмінною частотою. Тобто на ділянці між двома сусідніми екстремумами. На такій ділянці працює залежність:

$$f = \frac{F}{2 \cdot N}, \quad (3)$$

де f – частота сигналу на означеній ділянці, N – число вибірок, F – частота дискретизації.

Наступний етап дослідження – визначення частот нот хроматичного звукоряду в частотних межах мел-шкали. Ноти хроматичного ряду йдуть із кроком у півтона, а частота кожної наступної ноти обчислюється за формулою:

$$f_n = f_0 \cdot 2^{n/12}. \quad (4)$$

Як відомо, діапазон частот, які використовує людина у мовному сигналі (мовне ядро) 0,3–3,4 кГц. Мовою музичної теорії 4–7 октави у стандартній міжнародній системі нумерації октав (в радянській нумерації це 1–4 октави). Чотири октави мають 48 хроматичних нот. Але насправді в діапазон мовного ядра входить 42 ноти. Сигнал з частотою дискретизації 44,1 кГц передаючи найвищу ноту в межах мовного ядра А7 (3520 Гц), яку сприймає людське вухо, за довжину хвилі має 0.000284с, тобто 6 вибірок у напівперіоді. Якщо напівперіод, це свідчить, що це не мовний сигнал. На етапі апроксимації такий напівперіод потрібно або вилучити, або збільшити до 6 вибірок. Таким чином для кодування нормалізованих за хроматичним звукорядом частот мовного ядра достатньо літер латинської абетки у малому і великому регістрі.

Експериментальні дослідження фонемної ідентифікації. У запропонованому підході нормалізація мовного сигналу здійснюється у часовій області шляхом уніфікації довжини хвиль між сусідніми екстремумами (максимумами та мінімумами амплітуди). Для кожної хвильової ділянки визначається кількість вибірок між екстремумами, після чого вона трансформується у нормалізовану ділянку із фіксованою кількістю вибірок, що відповідає еталонній частоті (наприклад, частоті конкретної ноти у музичному ладі). Така заміна виконується послідовно для всіх хвильових елементів сигналу.

У результаті реальний сигнал може подовжитись або скоротитись на випадкове число вибірок, але формальна структура хвиль залишається консистентною та узгодженою із вибраною системою

еталонів. Тим часом дослідження фонемної ідентифікації у мовному сигналі, модифікованому шляхом нормалізації у часовій області, засвідчило збереження задовільного рівня розпізнаваності. Нормалізація значень амплітуди має вплив на якість ідентифікації звуку ще менший ніж нормалізаційна уніфікація довжин хвиль. На діаграмах подається первинний сигнал і сигнал модифікований частотною і амплітудною нормалізацією (рис. 1 та 2).

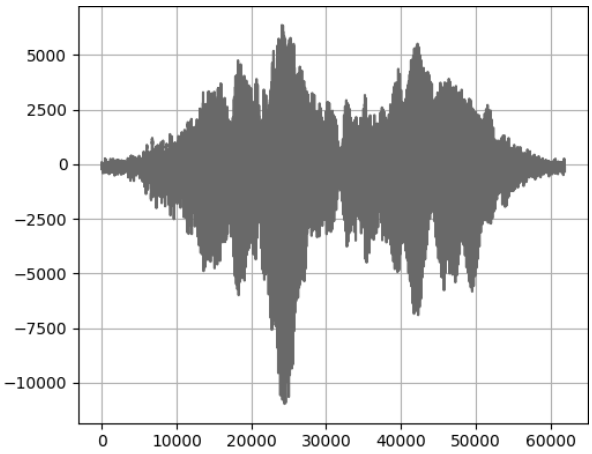

РИС. 1. Осцилограма вхідного сигналу

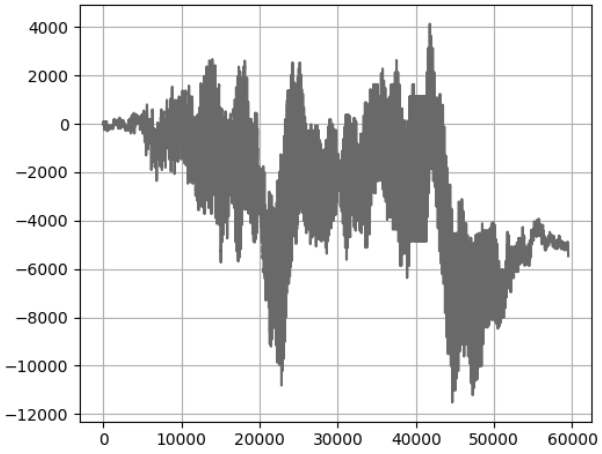

РИС. 2. Частотно нормалізований сигнал

Як видно з діаграм модифікований сигнал у часовій області змінився не критично. Слуховий апарат людини не ідентифікує трансформації. Стосовно модифікацій у амплітудній області, їх також дуже складно ідентифікувати слуховим апаратом.

Тим не менше створення способів частотної і амплітудної нормалізації потребують вдосконалення і на сьогодні відбувається створення таких способів. На діаграмі представлено нецентрований сигнал, отриманий шляхом амплітудної нормалізації. Необхідність приведення сигналу до центрованого вигляду може бути необов'язковим з огляду на подальші підходи оброблення мовного сигналу. Для методів, які використовують акустичні ознаки, центрування це важливий елемент дослідження, тому що нецентрованість сигналу впливає на спектральну оцінку сигналу. Тим часом використання апріорних значеннь частоти не потребує додаткових спектральних оцінок. Однак такі методи центрування нормалізованого сигналу можуть бути використані поряд з новими методами амплітудної нормалізації. Приклад такого методу центрування це метод, результат роботи якого подано на діаграмі рис. 3. В будь-якому разі перераховані методи потребують вдосконалення.

Формування з мовного сигналу навчальної послідовності для нейромережі. У запропонованій моделі мовлення розглядатиметься не лише як сигнал, а як структурована послідовність знакових одиниць. Кожна одиниця репрезентує диференційовану зміну акустичних параметрів – миттєвої частоти (періоду коливань) та амплітуди – між двома сусідніми часовими моментами. Ці зміни відповідним чином нормалізуються та кодуються у вигляді дволітерного символьного представлення XY, де:

X – літера, що відповідає нормалізованій різниці миттєвих довжин хвиль ($\Delta\lambda$),

Y – літера, що відповідає нормалізованій різниці амплітуд (ΔA).

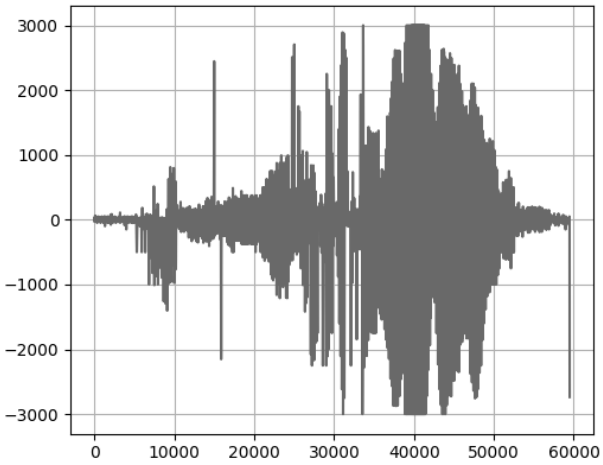

РИС. 3. Центрований сигнал

Масив символів побудовано з обмеженого алфавіту з 52 літер ($a-z, A-Z$) для кожної координати, що дозволяє утворити 676 унікальних пар XY – потенційних елементарних знакових одиниць.

Таким чином, безперервний мовний сигнал подається як дискретна послідовність знакових пар: $aH, bI, cG, aH, AI, \ldots$.

Ця послідовність це основа для виявлення фонемних шаблонів, які будуються як сукупності характерних підпослідовностей таких пар.

Отже на основі унікальних пар XY побудовано словник, що зіставляє кожну пару з цілим числом у межах [0; 675]. Цей словник використовується на рівні входу в модель, яка навчається на числових індексах, а не на символьних комбінаціях.

При цьому дотримується концепція, за якою одна й та сама літера X у позиції частотної координати не має однакового значення з тією самою літерою X у позиції амплітудної координати, що відображає їхню семантичну роздільність у структурі ознаки.

Мовний корпус анотовується на рівні фонем, з чітким визначенням меж кожної фонеми. Кожна фонема відповідає вікну послідовності з k знакових одиниць (k = 3…7), обраних емпірично. Таким чином, формується набір пар:

$[aH, bI, ZG, aH, aH, \ldots] \rightarrow$ фонема «а»

$[AI, aH, BI, AH, Ai, \ldots] \rightarrow$ фонема «о»

Кожна навчальна вибірка – це:

вхід: послідовність з k індексованих XY-пар,

вихід: фонемна мітка.

Зібрані знакові послідовності групуються за фонемами. Для кожної фонеми аналізуються найбільш типові послідовності пар, середньостатистична «форма» шаблону та варіативність (кластеризація всередині фонеми). Такий підхід корелюється з підходами запропонованими в роботі [11].

Висновки. Подані вище дослідження надають підстави говорити про можливість розглядати фонему, зокрема, мовний сигнал взагалі, як знакову послідовність нормалізованих параметрів – частоти та амплітуди. В такій послідовності кожен елемент виступає у ролі символу, що фіксує сталі відношення між акустичними характеристиками та їхнім місцем у структурі мовної ідентифікації. Тому є підстави стверджувати, що формалізація мовного сигналу як послідовності знакових

одиниць, які відповідають за диференційні апріорні ознаки мовного сигналу, що дозволяють побудувати знакову фонемну модель, це не що інше, як семіотична репрезентація.

Таким чином, запропонований підхід до моделювання мовного сигналу ґрунтується на принципах семіотики – це наука про знаки та знакові системи. Його головна перевага полягає у тому, що кожна одиниця аналізу (XY) – семіотична цілісність, яка має форму (вираження – акустичний змінний вектор) і функцію (зміст – відповідність фонемі або частині фонеми).

Це надає кілька ключових переваг:

- пояснюваність моделі: на відміну від «чорних скринь» класичних моделей, семіотична структура дає змогу виявити конкретні елементи, що формують фонему;
- інтерпретованість шаблонів: можна простежити, які послідовності семіотичних одиниць відповідають тій чи іншій фонемі, а також які з них – ключові;
- можливість узагальнення: навіть при варіативному сигналі система здатна виявляти структурні схеми через кластери, що відповідають алофонам або фонетичним умовам;
- гнучкість до низькоресурсних мов: семіотична модель не потребує величезних обсягів транскрибованих даних і може бути адаптована на базі правил або часткової анотації.

Традиційні системи розпізнавання мовлення (як CTC, трансформери, HMM) оперують або з акустичними ознаками без структурного представлення, або з графемними/фонемними транскрипціями, які не мають зв’язку з внутрішніми змінами у сигналі. Тоді як запропонований підхід не просто формалізує сигнал як вхід до моделі, а концептуалізує його як послідовність знаків, що дозволяє поєднати рівні сигналу і лінгвістичної структури, будувати правила на базі мовних категорій (наприклад, інтонаційні зміни, довжина тощо), уніфікувати вхідні ознаки для різних мов.

Тому що підхід до побудови фонемної моделі мовного сигналу, що базується на дискретному представленні акустичних змін у вигляді семіотичних одиниць XY , дозволяє не просто аналізувати сигнал, а структурувати його як послідовність знакових елементів з чіткою внутрішньою формою та функцією. На основі такого підходу побудовано семіотичну репрезентацію мовлення, яка забезпечує не лише ефективне розпізнавання, а й високу інтерпретованість мовного сигналу. Також розроблення рекурентної нейронної моделі створює можливість точного відтворення фонем на основі шаблонів із семіотичних одиниць;

Одна з найважливіших переваг запропонованого підходу – його ефективність у мовах із обмеженою кількістю навчальних ресурсів. Оскільки семіотичні одиниці базуються не на мові, а на універсальних фізичних змінних сигналу, можливо:

- швидко створювати фонемні шаблони для нових мов;
- застосовувати transfer learning між різними мовами за допомогою узагальнених семіотичних структур;
- зменшити потребу у великій кількості вручну транскрибованих даних, що є серйозною перешкодою для багатьох регіональних або корінних мов.

Семіотичний шар також полегшує використання правил або мовних гіпотез у системі розпізнавання, що додатково зменшує залежність від обсягів корпусів.

Моделювання запропонованих процесів відбувалося у середовищі програмний каркас FrameSound [12], що функціонує на базі Python 3.10 та IDE PyCharm 2023.3.2, а також Jupyter Notebook з використанням суперкомп’ютерного кластерного комплексу (СКІТ) Інституту кібернетики імені В.М. Глушкова НАН України і бібліотеки TensorFlow.

Авторські внески: Безвербний І.А. – концептуалізація, моделювання процесів, програмне забезпечення; Сосненко К.П. – створення і оброблення зразків мовлення, узагальнення, написання, редагування.

Наявність даних: на сайті https://b38xktoyq8.execute-api.us-east-1.amazonaws.com/prod подано окремі розробки на основі запропонованих ідей.

Список літератури

1. Rabiner L.R., Juang B.H. Fundamentals of Speech Recognition. Englewood Cliffs, NJ : Prentice-Hall, 1993. 507 p.
2. Graves A., Mohamed A., Hinton G. Speech recognition with deep recurrent neural networks. Proceedings of IEEE *International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2013*. P. 6645–6649. https://doi/org/10.1109/ICASSP.2013.6638947
3. Graves A., Fernández S., Gomez F., Schmidhuber J. Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks. Proceedings of the *23rd International Conference on Machine Learning (ICML)*, 2006. P. 369–376. https://doi/org/10.1145/1143844.1143891
4. Baevski A., Zhou Y., Mohamed A., Auli M. Wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. *Advances in Neural Information Processing Systems*. 2020. Vol. 33. P. 12449–12460.
5. Horndasch A., Noeth E., Batliner A., Warnke V. Phoneme-to-grapheme mapping for spoken inquiries to the semantic. Isca-archive INTERSPEECH, 2006. ICSLP. https://www.isca-archive.org/interspeech_2006/horndasch06_interspeech.pdf (звернення: 17.09.2025)
6. Sazhok M.M., Robeiko V.V., Smoliakov Ye.A., Zabolotko T.O., Seliukh R.A., Fedoryn D.Ya, Yukhymenko O.A. Modeling domain openness in speech information technologies. *Control systems and computers*ю 2023. No. 4. P. 19–26.
7. Semotiuk M.V., Palagin A.V. Technocratic model of the human auditory system. arXiv preprint arXiv:2310.05639, 2023. https://arxiv.org/abs/2310.05639 (звернення: 17.09.2025)
8. Семотюк М.В., Безвербний І.А. Адаптивний алгоритм виділення фонем у мовному сигналі. *Комп'ютерні засоби, мережі та системи*. 2017. № 16. С. 14–19. http://jnas.nbuv.gov.ua/article/UJRN-0000848988
9. Безвербний І.А. До питання виділення фонем у мовному сигналі за допомогою ефекту стоячої хвилі. *Комп'ютерні засоби, мережі та системи*. 2019. № 18. С. 32–35. http://jnas.nbuv.gov.ua/article/UJRN-0001084065
10. Безвербний І.А. Чірплет-аналіз мовних сигналів на основі перетворення Гільберта – Хуанга. *Кібернетика та комп'ютерні технології*. 2025. № 1. С. 74–80. https://doi.org/10.34229/2707-451X.25.1.7
11. Груша В.М. Інтелектуальна обробка даних від хлорофіл-флуорометричних сенсорів. *Кібернетика та комп'ютерні технології*. 2022. № 1. С. 42–48. https://doi.org/10.34229/2707-451X.22.1.5
12. Свідоцтво про реєстацію авторського права на твір №110368.

Одержано 17.09.2025

Безвербний Ігор Анатолійович,
кандидат технічних наук, старший науковий співробітник
Інституту кібернетики імені В.М. Глушкова НАН України, Київ,
https://orcid.org/0009-0005-4344-3068
ihorbezverbnyi@gmail.com

Сосненко Катерина Петрівна,
молодший науковий співробітник
Інституту кібернетики імені В.М. Глушкова НАН України, Київ.
https://orcid.org/0000-0002-3202-2680

UDC 81’1:003:004.932.2(045)

Ihor Bezverbnyi *, Kateryna Sosnenko

Semiotic Approach to the Construction of a Phoneme Model of the Speech Signal

V.M. Glushkov Institute of Cybernetics, NAS of Ukraine, Kyiv
** Correspondence: ihorbezverbnyi@gmail.com*

Introduction. The speech signal is characterized by high variability of its physical parameters; however, phonemes retain stability of identification even under significant fluctuations of frequency and amplitude. This provides a basis for constructing models that abstract from precise acoustic values and rely on the functional and semiotic nature of speech. Such an approach enables semiotic representation of the signal, where relative parameter changes play a key role. Its methodological foundation is associated with the idea of speech signal

normalization by the frequencies of the twelve-tone chromatic scale, which in turn finds confirmation in the psychoacoustic properties of human hearing and the anatomical structure of the human hearing apparatus.

The purpose of the study is to form a semiotic dictionary of the speech signal based on the calculation of frequencies required for speech transmission, normalized according to the twelve-tone chromatic scale, and to further develop a semiotic model that ensures the possibility of building interpretable speech recognition systems.

Results. The study substantiates the use of the concept of correlation between speech signal frequencies and the twelve-tone chromatic scale as the ideological basis of normalization. It is proposed to encode the signal through pairs of normalized conjugations of the frequency difference and the amplitude difference. Such representation creates a system of sign structures with a clear internal form and function, allowing not only signal analysis but also its interpretation. Based on this approach, a semiotic representation of speech has been constructed, which provides not only effective recognition but also a high degree of interpretability of the signal. In addition, the development of a recurrent neural model creates the possibility of accurate phoneme reproduction on the basis of semiotic unit patterns, opening prospects for further integration of the semiotic approach with deep learning methods.

Conclusions. Semiotic representation of the speech signal in the form of discrete sign units opens perspectives for the creation of interpretable automatic speech recognition systems. The proposed model combines theoretical novelty with practical significance, contributing to the development of computational linguistics and artificial intelligence technologies.

Keywords: speech signal, linguistic structure, phoneme model, interpretability, semiotic representation, recurrent neural model.