УДК 519.688

О. П. Луценко, С. Ф. Сірик, Л. В. Мащенко, Г. А. Омельницький

Дніпровський національний університет імені Олеся Гончара

СТРУКТУРА ІНФОРМАЦІЙНОЇ ТЕХНОЛОГІЇ ВИЯВЛЕННЯ ДЕФЕКТІВ МОВЛЕННЯ В МОВНОМУ СИГНАЛІ

Описано запропоновану авторами інформаційну технологію виявлення дефектів мовлення, методи фонемної сегментації мовного сигналу, класи дефектів, що виявляються, а також способи виявлення цих дефектів. Надано проектно-системну документацію програмного пакета, в якому реалізовано описані алгоритми.

Ключові слова*: мовний сигнал, дефекти мовлення, фонема, сегментація, інформаційна технологія.*

Описана предложенная авторами информационная технология обнаружения дефектов речи, методы фонемной сегментации речевого сигнала, классы дефектов, а также способы выявления этих дефектов. Представлена проектно-системная документация программного пакета, в котором реализованы описанные алгоритмы.

Ключевые слова: *речевой сигнал, дефекты речи, фонема, сегментация, информационная технология.*

The purpose of this paper is to describe the methods and algorithms of information technology proposed by authors for speech defects detection. The task of speech defects detection contains the tasks of segmentation of the speech signal, phoneme classification, as well as comparing the possibly impaired speech with a properly spoken reference signal. Comparisons should be made both for the signal as a whole (detection of missed phonemes) and for each phoneme separately (detection of defective phonemes). The segmentation algorithm used in information technology is based on the assertion that, at interphonemic transitions, the signal undergoes significant changes on many scales at once and is characterized by an increase in wavelet detailing coefficients, while in stationary areas the wavelet coefficients are grouped near certain scales. Phoneme wise presentation of the signal makes it possible to identify defects related to the incorrect pronunciation of the phoneme present in the word - the case where the total number of phonemes in the word remains unchanged. It is also possible to detect complete loss of the phoneme or the appearance of extra phonemes. To map phonemes to references, we use a dynamic timeline transformation algorithm, the vector of parameters of which

is the Euclidean distance between the Fourier decomposition bands. In order to determine the degree of proximity of the input to the reference, an additional limitation is proposed, according to which the distance from the input signal to the reference signal should be significantly different from the distances for other files. As a response threshold, 1/3 of the variance of distances between files was used. When the number of phonemes in a word does not match the reference, each phoneme of the reference is mapped to an input file, so that one of the input phonemes is matched, without violating the order of phonemes. If more than one non-identified section in a row is detected during segmentation, the sections are combined and compared to the base of the reference phonemes as one phoneme. If the base of the reference phonemes does not match the nearest result to the adjacent left or right phonemes, the segment of the speech is marked as defective. The described algorithms are implemented as part of information technology for assisting in the formulation of the pronunciation by providing the ability to independently assess the quality of the pronunciation.

Keywords: *speech signal, speech defect, phoneme, segmentation, information technology.*

Вступ. Для виявлення та лікування вад мовлення важливим є аналіз вимови пацієнта. Якщо для виявлення вад досить однієї перевірки у лікаря, то для лікування одного найпростішого дефекту, наприклад, простої дислалії однієї групи звуків, знадобиться у середньому 10 занять зі спеціалістом, не враховуючи занять вдома. При самостійних заняттях важливою є можливість самостійного контролю свого прогресу, який було б зручно здійснювати із застосуванням комп'ютерної програми. Розробка програмного забезпечення автоматичного виявлення відхилень у вимові є актуальною.

Аналіз літературних даних. Задача виявлення дефектів мовлення містить підзадачі сегментації мовного сигналу, класифікації фонем, а також порівняння сигналу з імовірним дефектом з правильно вимовленим еталонним сигналом. Порівняння необхідно виконувати як для сигналу в цілому (виявлення пропущених або зайвих фонем), так і для кожної фонеми окремо (виявлення фонем, що вимовляються з дефектами). Задача сегментації регулярно розглядається у працях вітчизняних і зарубіжних вчених [1–4]. В якості методу порівняння з еталонами можуть виступати алгоритми класифікації за вектором ознак, наприклад алгоритм динамічної трансформації часової шкали [5]. Виявлення дефектів мовлення досліджувалося у численних наукових працях у галузі медицини, в яких було розглянуто фізіологічні причини виникнення дефектів, однак у медичних працях у силу приналежності до іншої галузі наук бракує формального математичного опису мовних дефектів. Дослідження з розробки

автоматизованих систем розпізнавання дефектів мовлення або вивчають конкретний дефект [6], або присвячені проблемі покращення розпізнавання сигналу з дефектами і не розглядають класифікацію самих дефектів [7]. Таким чином, задача розробки інформаційної технології автоматизованого виявлення і класифікації дефектів мовлення залишається не вирішеною.

Постановка цілей статті. Мета даної статті – опис методів та алгоритмів запропонованої авторами інформаційної технології розпізнавання дефектів мовлення. Описано методи фонемної сегментації мовного сигналу, класи дефектів, що виявляються, а також способи виявлення цих дефектів.

Структура інформаційної технології. Виявлення дефектів мовлення є багатоетапною задачею, яка потребує багатокомпонентної інформаційної технології. Технологія повинна вирішувати задачі первинної обробки даних, фонемної сегментації та набір алгоритмів виявлення дефектів. На рис. 1 наведено діаграму варіантів використання інформаційної технології.

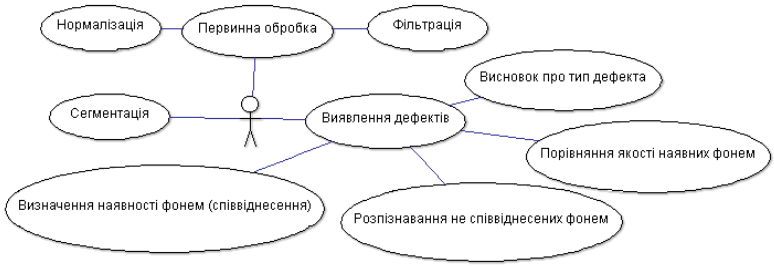

Рисунок 1 – Діаграма варіантів використання інформаційної технології

Ключовим завданням для рішення задачі виявлення дефектів мовлення є сегментація мови, бажано відповідно до фонетичної транскрипції мови. У процесі розпізнавання необхідно спочатку сегментувати мовний сигнал на характерні елементи, визначити тип сегмента, а потім проводити порівняння ідентичності сигналів за наявністю сегментів, а також за подібністю вимови окремих сегментів.

Ручна сегментація вимагає значних витрат сил і часу. Крім того, практично неможливо відтворити результати ручної сегментації внаслідок мінливості людського зорового і слухового сприйняття.

Автоматична сегментація не безпомилкова, проте вона несуперечлива за своєю суттю, і її результати відтворювані.

На відміну від сегментації слів, алгоритми для якої засновані на пошуку паузи між словами і є досить простими, сегментація фонем ускладнюється відсутністю значимих пауз. Сегментація фонем проводиться шляхом визначення моментів змін у частотних характеристиках сигналу.

Алгоритм сегментації, використаний в пропонованій інформаційній технології, було описано у статті [4]. Він заснований на частотному розкладанні з використанням вейвлетів. В основі алгоритму лежить твердження, що на міжфонемних переходах сигнал зазнає значних змін відразу на багатьох масштабах дослідження і, відповідно, характеризується значним зростанням вейвлет-коефіцієнтів для багатьох рівнів деталізації, в той час як на стаціонарних ділянках фонем вейвлет-коефіцієнти виявляються згрупованими поблизу певних масштабів. Відшукання міжфонемних переходів зведено до відшукання моментів змін вейвлет-коефіцієнтів на окремих рівнях розкладання.

Розіб'ємо сигнал на кадри довжиною 256 відліків (приблизно 10 мс). При цьому кадри розташуємо з половинним перекриттям так, щоб перші 128 відліків кожного кадру співпадали з останніми 128 відліками попереднього кадру. Для зменшення спектрального розтікання до кожного кадру застосуємо віконну функцію Хеммінга. До кожного з вікон проведемо вейвлет-розкладання.

Вейвлет-розкладання сигналу $S(t)$ позначимо як:

$$S(t) = \sum_{k=0}^{N/2^n-1} s_{nk}\varphi_{nk} + \sum_{j=1}^{N}\sum_{k=0}^{N/2^n-1} d_{jk}\psi_{jk},$$

$$\varphi_{nk} = 2^{n/2}\varphi(2^n t - k), \partial e\ j,k \in \mathrm{Z},$$

$$\psi_{nk} = 2^{j/2}\psi(2^j t - k), \partial e\ j,k \in \mathrm{Z}.$$

де n – кількість рівнів декомпозиції, s_{nk}, d_{jk} – коефіцієнти апроксимації та деталізації вейвлет-розкладання, φ – масштабна функція вейвлета, ψ – базисний вейвлет.

Для кожного вікна на кожному рівні декомпозиції знаходимо енергію коефіцієнтів деталізації:

$$E_n(i) = \sum_{j=1}^{2^n-1} d^2_{n,j\,2n-1_i}, \quad i = 0,...,2^{-M}N-1.$$

Знайдений ряд енергій згладжується трійками значень за правилом:

$$E'_n(i) = E'_n(i+1) = E'_n(i+2) = \max(E_n(i), E_n(i+1), E_n(i+2)) .$$

Міжфонемний перехід характеризується різкою зміною значення енергії на одному або декількох рівнях розкладання (авторами статті [4] запропоновано проводити 6 рівнів розкладання). Різкий перехід визначається як такий перехід, при якому перша похідна ряду за величиною порівнювана зі значенням ряду, при цьому значення енергії більше певного встановленого бар'єра E_{min} (низькі значення енергії еквівалентні тиші в оброблюваному частотному діапазоні):

$$h < | E'_n(i) - R_n(i) |,$$

$$h > | E'_n(i-1) - R_n(i-1) |, \text{ або } h > | E'_n(i+1) - R_n(i+1) |,$$

$$E'_n(i) > E_{\min} .$$

Приклад результатів роботи алгоритму наведено на рис. 2.

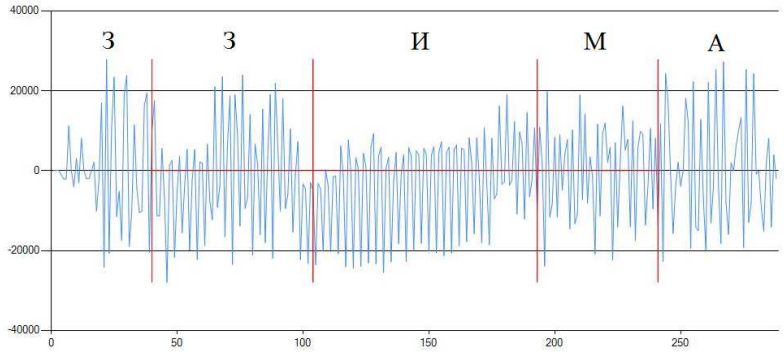

Рисунок 2 – Межі сегментів слова «зима», побудовані на графіку коефіцієнтів апроксимації 6 рівня вейвлет-розкладання

Як можна бачити з рис. 2, автоматична сегментація не позбавлена недоліків: фонемі |з| у прикладі відповідають дві ділянки, в чому можна переконатися, прослухавши файл і вручну зіставивши відносну довжину фонем. Однак, використання однакового алгоритму сегментації з однаковими налаштуваннями зазвичай породжує неточності одного характеру як в еталоні, так і у вхідному файлі. Таким чином, пофрагментне зіставлення файлів залишається ефективним.

Фонемна сегментація сигналу дає можливість працювати у двох напрямах:

1) виявляти дефекти, пов'язані з неправильною вимовою фонеми, яка присутня у слові – випадок, коли загальна кількість фонем у слові залишається незмінною;

2) виявляти повне випадання фонеми або появу зайвої фонеми. Слід зазначити, що, окрім випадку, коли фонема дійсно відсутня (так, при нечіткій вимові |p| у словах з двома приголосними підряд, наприклад, «грім», результат може сприйматися на слух як повне випадання фонеми |p|), випадання фонеми при автоматичній сегментації також може бути зумовлене дефектом сегментації внаслідок недостатньо чіткої вимови і злиття фонем.

Для співставлення фрагментів використовується алгоритм динамічної трансформації часової шкали, в якості вектора параметрів якого взято евклідові відстані між смуговими представленнями розкладання Фур'є.

Нехай маємо сигнал зі спектрально-часовим представленням $S(\omega_{kl})$, $k=\overline{1,128}$ $l=\overline{1,m}$, де m – кількість кадрів. Побудуємо спектрально-смугове представлення в 9 смугах з коефіцієнтів спектрального розкладання за правилом:

$$B_{1,l}=\sqrt{S^2(\omega_{1,l})+S^2(\omega_{2,l})},\ B_{2,l}=\sqrt{S^2(\omega_{3,l})+S^2(\omega_{4,l})},$$

$$B_{3,l}=\sqrt{S^2(\omega_{5,l})+S^2(\omega_{6,l})},\ B_{4,l}=\sqrt{S^2(\omega_{7,l})+S^2(\omega_{8,l})},$$

$$B_{5,l}=\sqrt{S^2(\omega_{9,l})+S^2(\omega_{10,l})},\ B_{6,l}=\sqrt{S^2(\omega_{11,l})+...+S^2(\omega_{15,l})},$$

$$B_{7,l}=\sqrt{S^2(\omega_{16,l})+...+S^2(\omega_{25,l})},\ B_{8,l}=\sqrt{S^2(\omega_{26,l})+...+S^2(\omega_{50,l})},$$

$$B_{9,l}=\sqrt{S^2(\omega_{51,l})+...+S^2(\omega_{110,l})}.$$

Сигнали $B_{1,l}-B_{9,l}$ зашумлені, і для кожного з них необхідно виконати фільтрацію на основі простого низькочастотного фільтра вигляду:

$$y_n=\sum_{k=-N}^{N}a_kB_{n-k},$$

$$a_k=\frac{1}{k\pi}\sin 2\pi kf_c,$$

$$a_0=2f_c,\ 0<f_c\le 0.5.$$

Матриця евклідових відстаней між смуговими представленнями двох сигналів будується за правилом:

$$d_{ik} = \left\{ \sqrt{\sum_{j=1}^{9} (y_{i,j,1} - y_{k,j,2})^2} \right\}, \; i = \overline{1, m_1}, \; k = \overline{1, m_2} \; ,$$

де m_1, m_2 – кількість кадрів у першому і другому файлах.

Знайдена матриця є вхідною матрицею для алгоритму динамічної трансформації часової шкали (DTW).

Слід враховувати, що алгоритм DTW ідентифікує вхідний сигнал за мінімальною відстанню до еталона, незважаючи на величину відстані як такої. Таким чином, певний еталон завжди буде поставлений у відповідність вхідному файлу. Однак може мати місце ситуація, коли еталон, який відповідає вхідному файлу, у вибірці не присутній взагалі, і обрано найбільш схожий еталон з доступних. Ця проблема гостро стоїть на першому етапі розпізнавання, так як вибірка еталонів на цьому етапі обмежена фонемами, які є в еталонній вимові. З метою визначення ступеня близькості входу до еталона запропоновано додаткове обмеження, згідно з яким відстань від вхідного сигналу до еталона повинна значно відрізнятися від відстаней для інших файлів. Поняття «значності» зв'яжемо з величиною дисперсії δ вибірки відстаней, поклавши в якості порога 1/3 дисперсії. Таким чином, правило прийняття рішення про відповідність виглядає як:

$$I = \arg\min_i (D_i),$$

$$D_i \geq \frac{\delta}{3},$$

де D – відстань між файлами, знайдена за допомогою алгоритму DTW.

При невиконанні правила фонема вважається неідентифікованою і ставиться у чергу на наступний етап обробки – порівняння з базовими фонемами. Неправильна вимова фонеми може виражатися в заміні однієї фонеми на іншу (|р| на |г| або |л|, |з| на |с|, твердої фонеми на м'яку тощо). Якщо фонему не вдалося зіставити з жодною фонемою в еталонній вимові, наступним етапом є зіставлення з базою фонем з метою виявлення можливого випадку заміни. У випадку, коли кількість фонем у слові не співпадає з еталоном, кожна фонема еталона зіставляється із вхідним файлом так, щоб поставити у відповідність одну з вхідних фонем, при цьому не порушуючи порядку слідування фонем. Якщо при сегментації виявлено більше однієї неідентифікованої ділянки поспіль, ділянки об'єднуються і порівнюються з базою еталонних фонем як одна фонема. Якщо за базою еталонних фонем найближчий результат не відповідає сусідній

зліва або справа фонемі, ділянка мови помічається як дефектна. Діаграму діяльності алгоритму виявлення дефектів наведено на рис. 3.

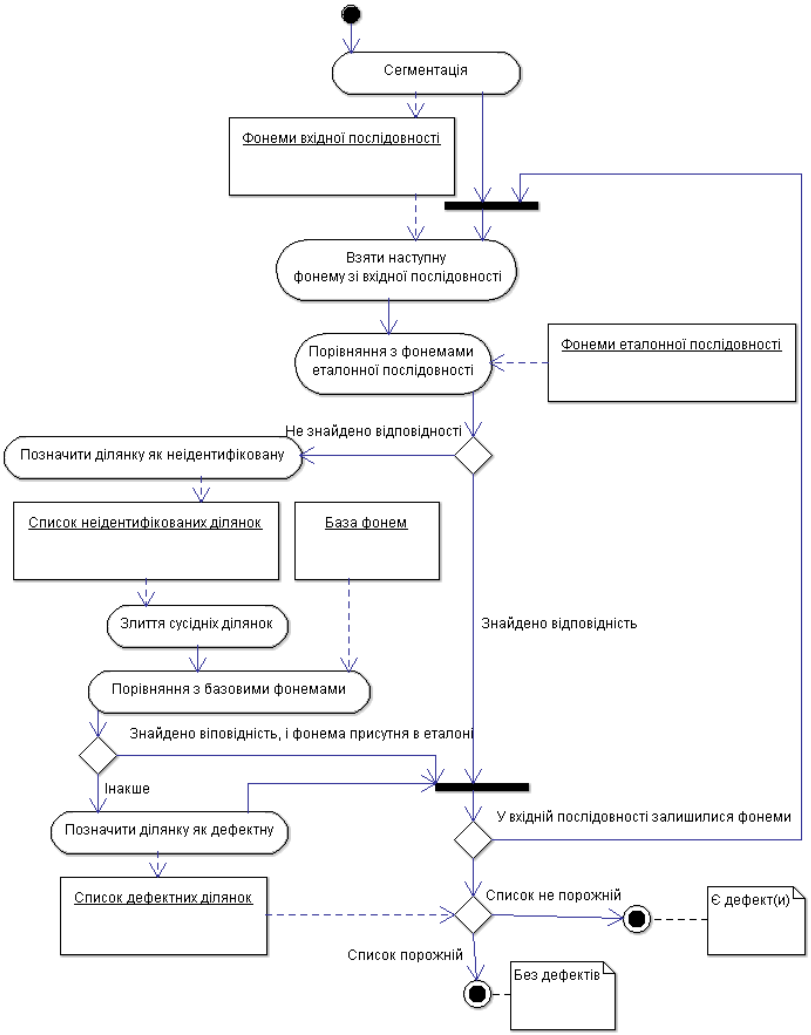

Рисунок 3 – Алгоритм виявлення дефектів

На останньому етапі необхідно визначити тип дефекту, маючи інформацію про спотворену ділянку. На цьому етапі відсутня необхідність розпізнавання фонеми, адже сказане слово або фраза, яка аналізується, апріорно відомі, також як і еталонне звучання. Потрібно лише виявити розташування фрагмента, який відсутній або спотворений у вхідному файлі.

Для визначення типу дефекту отриманий результат поєднується з апріорним знанням того, якій саме фонемі відповідає проблемна ділянка сигналу. Це дозволяє приймати рішення про клас дефекту [8]: сигматизм (дефект вимови свистячих і шиплячих звуків); ротацизм (дефект вимови звуків [Р], [Р ']); ламбдацизм (дефект вимови звуків [Л], [Л ']); каппацизм (дефект вимови звуків [К], [К ']); гаммацизм (дефект вимови звуків [Р], [Г ']); хітизм (дефект вимови звуків [Х], [Х ']); йотацизм (дефект вимови звуку [j]).

Висновки. Описані алгоритми реалізовано у складі інформаційної технології, призначеної для надання допомоги в постановці вимови шляхом надання можливості самостійного оцінювання якості вимови.

Напрям подальших досліджень: удосконалення алгоритмів автоматичної сегментації, автоматизація підбору налаштувань алгоритмів.

Бібліографічні посилання

1. Шелепов В.Ю., Ниценко А.В. Сегментация речевого сигнала, соответствующего заранее известному слову. *Искусственный интеллект*. 2014. № 4. С. 202–207.

2. Бурибаева А.К., Дорохина Г.В., Ниценко А.В., Шелепов В.Ю. Сегментация и дифонное распознавание речевых сигналов. *Труды СПИИРАН*. 2013. Вып. 8(31). С. 20–42.

3. Сорокин В.Н., Цыплихин А.И. Сегментация и распознавание гласных. *Информационные процессы*. Т. 4. № 2. С. 202–220.

4. Вишнякова О.А., Лавров Д.Н. Автоматическая сегментация речевого сигнала на базе дискретного вейвлет-преобразования. *Математические структуры и моделирование*. 2011. Вып. 23. С. 43–48.

5. Al-Naymat G., Chawla S., Taheri J.. Sparse DTW: A novel approach to speed up Dynamic Time Warping. URL: https://arxiv.org/pdf/1201.2969v1.pdf (дата звернення: 14.11.2019).

6. Velican V., Grigore O., Grigore C. Pattern Recognition Based Method Used in Identifying Impaired Speech. *Proceedings of the 2nd international*

conference on Applied informatics and computing theory. September 2011. P. 190–194.

7. Potamianos G., Neti C. Automatic Speechreading Of Impaired Speech. *Proceedings of the International Conference on Auditory-Visual Speech Processing*. Aalborg, 2001. P. 177–182.

8. Ganapathiraju A., Hamaker J., Picone J., Doddington G.R. and Ordowski M. Syllable-Based Large Vocabulary Continuous Speech Recognition. *IEEE Transactions on Speech and Audio Processing.* 2001. Vol. 9. N. 4. P. 358–366.

Надійшла до редколегії 14.11.2019.