УДК 682.142.68

Ю.В. Данюк

Харківський університет Повітряних Сил ім. І. Кожедуба, Харків

СЕГМЕНТАЦІЯ МОВНИХ СИГНАЛІВ З ВИКОРИСТАННЯМ ДИСКРЕТНОГО ВЕЙВЛЕТ-ПЕРЕТВОРЕННЯ

В статті запропоновано метод сегментації мовних сигналів на основі аналізу варіації рівня енергії вейвлет-спектру з розставленням кордонів на ділянках швидкої зміни енергії сигналу за всіма рівнями деталізації.

Ключові слова: *сегментація, мовний сигнал, дискретне вейвлет-перетворення.*

Вступ

Постановка проблеми. Одним з найважливіших завдань в системах обробки мови є завдання сегментації відповідно до фонетичної транскрипції мови. Для голосової верифікації характерні ознаки голосу повинні обчислюватися на певних сегментах мовного сигналу. Частота основного тону, характерна диктору, повинна обчислюватися на приголосних ділянках сигналу. Форма мовного тракту характеризується формантними частотами, які вимірюються на відомих голосних звуках, а швидкість артикуляції визначається за тривалістю перехідних процесів між артикуляторно-акустичними сегментами. [1] Сегментація необхідна при вирішенні зворотної задачі – відновлення форми мовного тракту з акустичного сигналу [2], яка може бути використана в наступних областях: системи стиснення і передачі мови в мобільній телефонії [3], синтезатори мови за довільним текстом [4], системи автоматичного розпізнавання мови.

У дослідницьких системах і на етапі попередньої розробки можливе використання ручної сегментації. Однак вона вимагає значних витрат сил і часу: по-перше, в злитій мові немає пауз між словами, по-друге, коартикуляція, яка виникає на кордоні звуків суттєво полегшуючи правильне сприйняття і розуміння мови ускладнює завдання пошуку меж сегментів. Крім того, практично неможливо точно відтворити результати ручної сегментації внаслідок суб'єктивності людського слухового і зорового сприйняття.

Подібних проблем не виникає при автоматичній сегментації, яка також не безпомилкова, але дає непогані результати.

Існує два основних типи алгоритмів сегментації мови. До першого типу відносяться алгоритми, які сегментують мову якщо відома послідовність фонем даної фрази. Інший тип алгоритмів не використовує апріорної інформації про фразу, і при цьому межі сегментів визначаються за ступенем зміни акустичних характеристик сигналу. При автоматичній сегментації бажано використовувати тільки загальні характеристики мовного сигналу, оскільки зазвичай на цьому етапі немає конкретної інформації про зміст мовного висловлювання.

Основний розділ

1. Сегментація з використанням кратномасштабного аналізу

Як відомо, мовний сигнал складається з квазістаціонарних ділянок, які відповідають приголосним і шиплячим фонемам, що перемежовуються ділянками з порівняно швидкими змінами спектральних характеристик сигналу (міжфонемні переходи, вибухові фонеми, переходи між словами) [5]. У межах стаціонарних ділянок значну роль для аналізу мовного сигналу відіграють спектральні особливості сигналу, що визначаються передавальною характеристикою мовного тракту і змінюється в процесі артикуляції. Можна сказати, що мовний сигнал характеризується нелінійними флуктуаціями різних масштабів. Тому досить ефективним для аналізу мовного сигналу представляється крупномасштабний аналіз і вейвлет-перетворення.

Вейвлет-розкладання мовного сигналу довжиною N відліків являє собою суму:

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

$$\varphi_{nk} = 2^{j/2}\varphi(2^j t - k), \text{ де } j,k \in Z$$

$$\psi_{jk} = 2^{j/2}\psi(2^j t - k), \text{ де } j,k \in Z,$$

де n – кількість рівнів декомпозиції; s_{nk}, d_{jk} – коефіцієнти апроксимації та деталізації вейвлет-розкладання; φ – скейлінг (масштабна) функція; ψ – базисний («материнський») вейвлет.

Оскільки вейвлет-коефіцієнти апроксимації відповідають передавальній характеристиці фільтра низьких частот, а вейвлет-коефіцієнти деталізації – високочастотному фільтру, то можемо розглядати поведінку мовного сигналу в різних частотних діапазонах.

Частотний діапазон нижче 125Гц не використовується, тому не містить інформації, важливої для задачі сегментації. Це обумовлено природою людської мови, що охоплює інтервал 150 – 4000 Гц. Таким чином, достатньо 6 рівнів розкладання (табл. 1).

Таблиця 1

Частотні діапазони

Рівень деталізації	Частотний діапазон Добеші 16, Гц	Частотний діапазон Мейера, Гц
рівень 1	2000-4000	2756-5512
рівень 2	1000-2000	1378-2756
рівень 3	500-1000	689-1378
рівень 4	250-500	345-689
рівень 5	125-250	172-345
рівень 6		86-172

2. Алгоритм сегментації

Сегментація мовного сигналу необхідна для виділення ділянок сигналу, які відповідають окремим структурним одиницям. Якщо розглядати фонеми як такі одиниці, то завдання сегментації зводиться до виявлення міжфонемних переходів. Вирішення цього завдання традиційними підходами досить проблематичне. Однак дискретне вейвлет-перетворення (DWT) дозволяє вирішити цю проблему для фонем, які відповідають порівняно протяжним квазістаціонарним ділянкам. Справа в тому, що на міжфонемних переходах сигнал зазнає значних змін відразу на багатьох масштабах дослідження і, відповідно, характеризується зростанням вейвлет-коефіцієнтів для багатьох рівнів деталізації, в той час як на стаціонарних ділянках фонем вейвлет-коефіцієнти виявляються згрупованими поблизу певних значень [7]. Таким чином , пошук міжфонемних кордонів може бути зведений до пошуку моментів збільшення вейвлет-коефіцієнтів на значній кількості рівнів масштабування. При цьому суттєвим є вибір вейвлетного базису, який повинен дозволяти описувати стаціонарний мовний сигнал з порівняно малим числом ненульових коефіцієнтів. Можливе використання декількох вейвлетних базисів для пошуку міжфонемних переходів у кожному з них з наступним об'єднанням результатів [6].

Для початку сигнал розбивається на ділянки, що перекриваються, до кожної з яких застосовується DWT. Для кожного фрейма та рівня декомпозиції n можна визначити енергію:

$$E_n(i)=\sum_{j=1}^{2^n-1} d^2_{n,j+2^{n-1}}, \text{ де } i=0,\ldots,2^{-M}N-1. \qquad (2)$$

Енергія сигналу (2) швидко змінюється від фрейма до фрейму для кожного рівня через шуми під час запису мовного сигналу . Для згладжування визначаємо $E`_n$, замінюючи значення E_n у вікні шириною 3 – 5 фреймів на максимальне значення E_{max} в цьому вікні. Для визначення швидкості зміни енергії обчислюємо похідну R. Міжфонемні переходи характеризуються невеликими, але швидкими змінами рівня енергії на одному або більше рівнях деталізації. Таким чином, критерієм вибору кордону фонеми має бути швидка зміна похідної при невисокому рівні енергії. Іншими словами, для кожного рівня деталізації шукаємо такі ділянки, на яких значення похідної близько за своїм абсолютним значення до рівня енергії на інтервалі, при цьому різниця не перевищує деякого граничного значення div_{opt}, а енергія на цьому інтервалі обов'язково більша ніж E_{min} як гарантія аналізу саме мовного сигналу, а не ділянки шуму:

$$div_{opt} \geq \left| |R_n(i)| - E`_n(i) \right|.$$

3. Удосконалення алгоритму

Потрібні деякі доробки алгоритму для більш точного визначення меж сегментів. Положення кордонів може різнитися між рівнями. Це пояснюється природою вейвлет-перетворення – розгляд сигналу на різних частотних діапазонах. Так для частини фонем тільки один з рівнів покаже значну зміну енергії, для інших – декілька.

Таким чином, на кожному рівні визначається тільки частина міжфонемних переходів. При цьому міжфонемний інтервал не може бути меншим порогового значення – мінімальної тривалості фонеми. Поріг встановлено в 25 мсек. Загальний алгоритм сегментації має такий порядок:

1. В якості попередньої обробки сигнал нормалізується: всі відліки діляться на максимальне значення, для встановлення єдиних порогових значень для будь-яких вхідних сигналів.

2. Вхідний сигнал розбивається на фрейми по 256 відліків при частоті дискретизації 16 кГц з перекриттям від 25 % до 50 %.

3. Кожен фрейм накривається вікном Хеммінга для усунення дефектів на краях.

4. До кожного фрейму застосовується вейвлет-перетворення з розкладанням до 6-го рівня декомпозиції.

5. Для кожного рівня декомпозиції визначається енергія, як сума квадратів значень коефіцієнтів деталізації E (2).

6. Оскільки енергія сильно змінюється від фрейму до фрейму через шуми, необхідне згладжування. Для цього обчислюється усереднена енергія $E`_n$ для кожного рівня декомпозиції, замінюючи значення енергії на максимальне E_{max} для кожних 3 на перших трьох, і на кожних 5 для наступних рівнів деталізації.

7. Для визначення швидкості зміни енергії обчислюється похідна R.

8. Критерії вибору кордонів фонем:

$$div_{opt} \geq \left| |R_n(i)| - E`_n(i) \right|;$$

$$div_{opt} < \left| |R_n(i+1)| - E`_n(i+1) \right|;$$

$$div_{opt} < \left| |R_n(i-1)| - E`_n(i-1) \right|;$$

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

9. Для об'єднання результатів розстановки кордонів між рівнями всі індекси об'єднуються в один вектор. Щоб уникнути помилкових кордонів, встановлюється мінімальний інтервал фонеми – 28 мсек.

4. Результати експерименту

Для експерименту використано 35 різних фрагментів, записаних при частоті дискретизації 16кГц.

На рис. 1 пунктирними лініями відзначені межі ручної розмітки, суцільними – автоматичної. Наведені графіки енергій і їх похідних для кожного рівня деталізації.

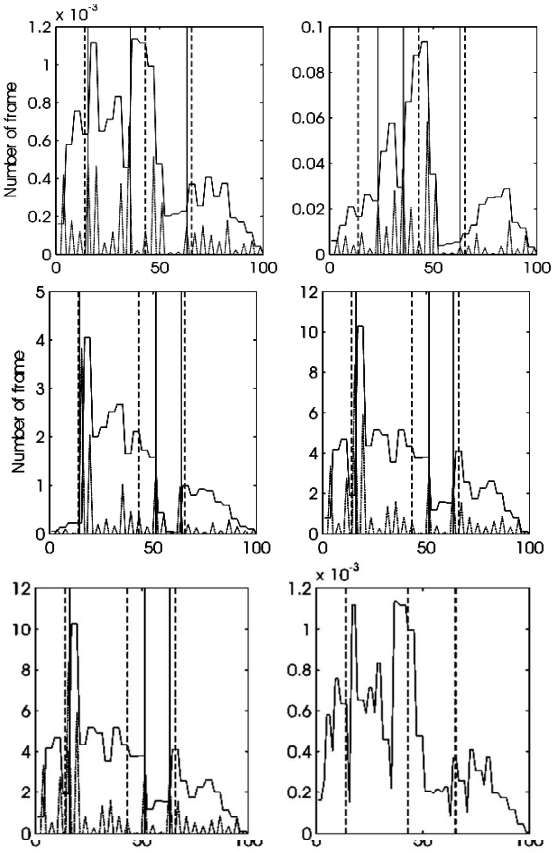

Рис. 1. Приклад сегментації слова “мама”

Експериментально оптимальними пороговими значення обрані $div_{opt} = 0.03$, $E_{min} = 0.005$. При цьому зі збільшенням порогового коефіцієнта зменшується чутливість алгоритму до змін мовного сигналу. Так, при значеннях 0.01 – 0.02 помітно виділення зайвих сегментів для приголосних, добре розділяються голосні звуки, які стоять поруч “оа”, “аі”. При великих значеннях порога кількість зайвих сегментів мала, але перестають розділятися голосові звуки. Результати експериментів показали незначну різницю в ефективності вейвлетів Майера, Добеші 16, Добеші 8 та Сімлета 6 порядку. Це говорить про можливість застосування всіх їх як базису розкладання з можливим майбутнім об'єднанням результатів для підвищення рівня розпізнавання кордонів сегментів.

Висновки

Запропонований метод сегментації заснований на дискретному вейвлет- перетворенні. Ефективність методу обумовлена природою мовного сигналу – зміни рівня енергії для ряду фонем виявляються тільки у вузькому діапазоні частот. Саме тому кордони найімовірніше детектувати, аналізуючи значення енергій піддіапазонів вейвлет-розкладання, а не сигналу в цілому, як у випадку перетворень, заснованих на Фур'є аналізі. В якості основного параметра визначення точної межі сегменту використовується швидкість зміни енергії при подальшому об'єднанні результатів розстановки кордонів між рівнями деталізації.

Список літератури

1. Рамишвили Г.С. Автоматическое опознавание говорящего по голосу / Г.С. Рамишвили. – М. : Радио и связь, 1981. – 224 с.

2. Макаров К.С. Построение и исследование артикуляторных кодовых книг для решения речевых обратных зада: дис. … на соиск. степ. к.т.н., ИППИ РАН, 2005. – 182 с.

3. Leonov A.S. Inverse problem for the vocal tract: identification of control forces from articulatory movements / A.S. Leonov, V.N. Sorokin // Pattern Recognition and Image Analysis. 2000. – V. 10, № 1. – P. 110-126.

4. Сорокин В.К. Синтез речи / В.К. Сорокин. – М. : Наука, 1992. – 392 с.

5. Сорокин В.Н. Сегментация и распознавание гласных / В.К. Сорокин, А,И. Цыплихин // Информационные процессы. – 2004. – Т. 4, № 2. – С. 202-220.

6. Ziolko B. Wavelet method of speech segmentation / B. Ziolko, S. Manandhar, R. Wilson, M. Ziolko // Proceedings of 14th European Signal Processing Conference EUSIPCO. 2006. – P. 23-35.

7. Ермоленко Т.В. Применение вейвлет-преобразования для обработки и распознавания речевых сигналов / Т.В. Ермоленко // Искусственный интеллект. – 2002. – № 4. – С. 200-208.

Надійшла до редколегії 6.11.2013

Рецензент: д-р техн. наук, проф. І.В. Рубан, Харківський університет Повітряних Сил, ім. І. Кожедуба, Харків.

СЕГМЕНТАЦИЯ РЕЧЕВЫХ СИГНАЛОВ С ИСПОЛЬЗОВАНИЕМ ДИСКРЕТНОГО ВЕЙВЛЕТ-ПРЕОБРАЗОВАНИЯ

Ю.В. Данюк

В статье предложен метод сегментации речевых сигналов на основе анализа вариации уровня энергии вейвлет-спектра с расстановкой границ на участках быстрого изменения энергии сигнала по всем уровням детализации.

Ключевые слова: *сегментация, речевой сигнал, дискретное вейвлет-преобразование.*

SPEECH RECOGNITION BASED ON THE METHOD OF SPECTRAL ESTIMATION

Y.V. Danyuk

The note proposed a method for segmentation of speech signals by analyzing the variation of the energy spectrum of the wavelet positioning boundaries in areas rapid changes in signal energy at all levels of detail.

Keywords: *segmentation, speech signal, discrete wavelet transform.*