УДК 621.3

к.т.н., доц.**Соловьев В.И.** (Вост.укр.НУ)
д.т.н., проф. **Рыбальский О.В.** (НАВД)

СПЕКТРАЛЬНЫЙ АНАЛИЗ И РЕЧЕВЫЕ ТЕХНОЛОГИИ

В статье рассмотрены результаты исследований авторов особенностей спектрального анализа сигналов речи на основе ортогонального и неортогонального время частотного преобразования. Подобный подход опирается на результаты современных исследований в области нейрофизиологии и психофизиологии слуха. Приведены результаты исследований по применению время-частотного преобразования на основе неортогонального базиса для анализа речевых файлов. Показано, что спектры на основе такого преобразования позволяют получать более достоверные оценки параметров речи в частотной области. Полученные результаты могут быть применены при построении систем идентификации диктора по параметрам его голоса и при построении систем автоматического распознавания речи. Кроме того, полученные результаты могут оказать существенное влияние на построение систем защиты речевой информации с ограниченным доступом.

Ключевые слова: частотный спектр, ортогональны базис, неортогональный базис, фонема, форманта, характеристики голоса.

Вступление. Современная теория речеобразования базируется на работах Гельмгольца [1]. Существенную детализацию и развитие эта теория получила фундаментальных трудах Фанта [2] и Фланагана [3]. Библиография в области речевых технологий в настоящее время насчитывает тысячи публикаций. Эти работы существенно детализируют основные моменты акустической теории речеобразования. Основополагающие труды второй половины 20-го века и последнего десятилетия в области нейрофизиологии и психофизиологии слуха внесли существенное понимание в механизмы нейрофизиологии восприятия звука.

Однако, несмотря на огромное число исследований в этой области, на сегодняшний день следует констатировать незавершенность акустической теории речеобразования. В настоящее время эта теория отвечает на большинство вопросов о механизмах формирования акустических колебаний при речеобразовании, но не дает ответ на принципиальный вопрос современной теории распознавания речи: каковы принципы кодирования информационной части речи? При этом имеется в виду не ответ на вопрос о механизмах кодирования. Эти механизмы на сегодняшний день общепризнанны. Вопрос заключается в конкретных информационных кодах для различных звуков речи, которые однозначно решают вопрос декодирования речевой информации, произносимой разными людьми. На сегодняшний день по-прежнему неясно, а корректна ли вообще постановка задачи о существовании таких кодов, например, в виде определенных функций спектров фрагментов речи.

Существование в настоящее время систем речевого ввода типа Speech Text и различных вариантов фонемических машин, по существу не отвечает на этот вопрос.

Как правило, системы первого типа (в существенной части и второго) построены на основе идеи нейронных сетей. Эти системы, в какой-то мере эмулируют возможный способ решения задач речевых технологий мозгом. Принято считать, что конгломераты нейронов мозга, обрабатывающие звуковую информацию, каким-то образом выделяют нечто общее в спектрах фрагментов речи, которые характеризуют информационную часть речи. При этом полагают, что мозг способен “находить” и “видеть” нечто общее в спектрах речевых фрагментов, что характеризуют, например, одну и ту же фонему при различных характеристиках речеобразующего тракта [4].

Использование выявленных закономерностей на сегодняшний день позволяет строить системы с определенной эффективностью распознающих звуки по спектральным характеристикам звуков речи. Однако, диапазон изменения этих характеристик таков, что системы в лучшем случае работают, лишь при вполне определенных характеристиках голоса.

При этом, одной из весьма существенных проблем в области обработки речевой информации является высокая вариабельность различных спектральных характеристик.

Так, например, из многочисленных исследований известно, что для гласных звуков речи [a], [i], [u], [e], [o] существуют определенные частотные диапазоны изменения формант [2; 3; 5–10]. Однако, даже для гласных звуков речи эти диапазоны, во-первых, пересекаются между собой, а во-вторых являются, все равно, среднестатистическими данными. Реальные частотные характеристики звуков на практике, вполне могут выходить за приводимые диапазоны.

Высокая степень изменчивости параметров спектра для одних и тех же выявляемых в исследованиях характеристик речи является серьезной проблемой при разработке общих концепций и теорий в различных областях речевых технологий.

Возникает она и при решении задач построения систем идентификации диктора по независимым от контекста характеристикам голоса, что требует эффективной обработки речевой информации.

Нет сомнения, что эта проблема существенным образом связана с природой акустической речевой информации.

Однако, возможна и иная причина весьма больших отличий параметров спектров, примеры которых приводятся в многочисленных исследованиях. Современный спектральный анализ в области речевых технологий базируется на кратковременном (“оконном”) преобразовании Фурье [11]. Неоднократно в различных исследованиях затрагивалась проблема как корректного применения спектрального анализа Фурье, так и возможностей эффективной физической интерпретации результатов спектрального анализа.

В данной статье представлены результаты исследований, которые уточняют ряд известных положений и могут ответить на ряд вопросов эффективного применения спектральных методов в речевых технологиях для малых временных интервалов.

Постановка задачи исследования. В приводимых ниже исследованиях будем опираться на ряд известных положений в области речевых технологий. Большинство этих положений являются следствием огромного числа экспериментальных исследований в различных смежных областях речевых технологий – нейрофизиологии, психофизиология восприятия речи, лингвистики и цифровых методов обработки аудиоинформации. По нашему мнению, в области речевых технологий, ввиду сложности объекта исследований, не существует однозначно трактуемых положений и результатов исследований. Но мы полагаем, что любое исследование должно быть учтено, поскольку этими исследованиями определяется достигнутое на сегодняшний день понимание вопросов речевосприятия.

Как указывалось выше, будем считать, что вся наиболее существенная информация, необходимая для распознавания речи, полностью содержится в параметрах звуковой волны фрагментов речи, представленной в цифровой записи.

В нашем исследовании рассматривается влияние на результаты спектрального анализа фрагментов речи в виде звуковой волны с точки зрения воздействия ряда факторов.

Первым существенным фактором приводимого ниже анализа является математическая “гладкость” спектров и описывающих их функций. Этот фактор весьма важен с точки зрения математического описания сигналов как для получения достоверных оценок спектров и их параметров, так и для построения эффективных алгоритмов локализации локальных экстремумов в частотной области. “Гладкость” спектров весьма существенно зависит от параметров окна преобразования и ряда параметров спектрального анализа.

Второй существенный фактор – достоверность физической интерпретации спектров и их функций, т.е. соответствие этой интерпретации, являющейся следствием вида конкретного спектра, физическому наличию частотных составляющих в исследуемом сигнале. Этот фактор также весьма существенно зависит от нескольких параметров спектрального анализа.

Наконец, весьма важным фактором является стабильность характеристик спектра и его локальных максимумов, связанных с формантами речи, как для фрагментов одной и той же фонемы, так и для характеристик голоса диктора.

Поставим задачу выявления факторов и параметров спектральных преобразований фрагментов речи, которые позволяют получить более “гладкие” и устойчивые спектральные функции и позволяют более корректную физическую интерпретацию спектральных функций.

Спектральный анализ фрагментов речи. Рассмотрим влияние фактора выбора параметра временного окна на результаты спектрального анализа характеристик звуковой волны.

Будем рассматривать фрагменты речи как дискретные временные ряды амплитуды звуковой волны $A(t_i)$. Будем рассматривать преобразование Фурье для различных малых временных интервалов фрагментов речи. Исследования в области нейрофизиологии и психологии восприятия звука показывают, что существенная часть обработки звуковой информации в частотной области нейронами слухового анализатора осуществляется на временных интервалах порядка 20–100 мс [2; 3; 10]. С другой стороны, известно, что большинство фонем речи могут быть “сконструированы” на основе мультифрактального подобия из малых составных частей фрагментов фонем. Это временные интервалы порядка 15–20 мс [11]. Мультифракталы в этих исследованиях трактуются на основе концепции Мандельброда [12–17]. Однако, классический спектральный анализ на малых временных интервалах (порядка 20мс) позволяет анализировать спектры с интервалом частоты дискретизации порядка 50 Гц.

В то же время, нейрофизиологические данные показывают, что в диапазоне частот до 500 Гц, слуховые анализаторы человека имеют частотное разрешение порядка 1 Гц [4].

Рассмотрим дискретное преобразование Фурье временного ряда амплитуд звуковой волны на малом временном интервале (20 мс) в следующем виде:

$$S(k) = abs\left(\sum_{t_{ij}=0}^{N_m} [\exp(-2\pi i / N)(t-1)(k-1)] A(t) \right), \qquad (1)$$

где $i = \sqrt{-1}$, abs – модуль комплексного преобразования, $S(k)$ - амплитуда спектра, $A(t)$ – амплитуда звуковой волны для дискретного отсчета *t*.

Параметр k в (1) будем рассматривать не как дискретную переменную номера частоты спектра, а как непрерывную переменную, которая может изменяться в диапазоне $1 \le k \le N/2$. Для дискретных значений *k* в этом диапазоне выражение (1) является обычным оконным дискретным преобразованием Фурье с ортогональным базисом. Для дробных значений *k* – это выражение является преобразованием с неортогональным базисом и произвольной частотой - $2\pi k / N$. Теория информации не исключает физического существования любых частотных составляющих в диапазоне частот от 1/T до частоты Найквиста (T – временное окно преобразования).

Проиллюстрируем наши предыдущие и последующие рассуждения рассмотрением двух спектров для одного и того фрагмента речи, показанных на рис. 1 и рис. 2.

Первый спектр (рис. 1) является оконным преобразованием Фурье с дискретным набором частот (ортогональный базис, дискретность шага по частоте – 50 Гц). Второй (рис. 2) – неортогональный базис с произвольным набором частот (дискретность шага по частоте – 1 Гц).

Для спектральных преобразований на малых временных интервалах точность локализации максимумов в первом случае не может превышать 50 Гц. Для неортогональных базисных функций в данном примере точность локализации максимумов спектра – порядка 1Гц.

Неудобство второго варианта преобразования заключается в невозможности применения быстрых алгоритмов спектрального анализа, близких по эффективности быстрому преобразованию Фурье. Однако, как показывают проведенные нами многочисленные исследования спектров на основе набора неортогональных базисных

функций с дробным шагом по частоте, эти спектры дают возможность получения устойчивых оценок важных частотных параметров речи.

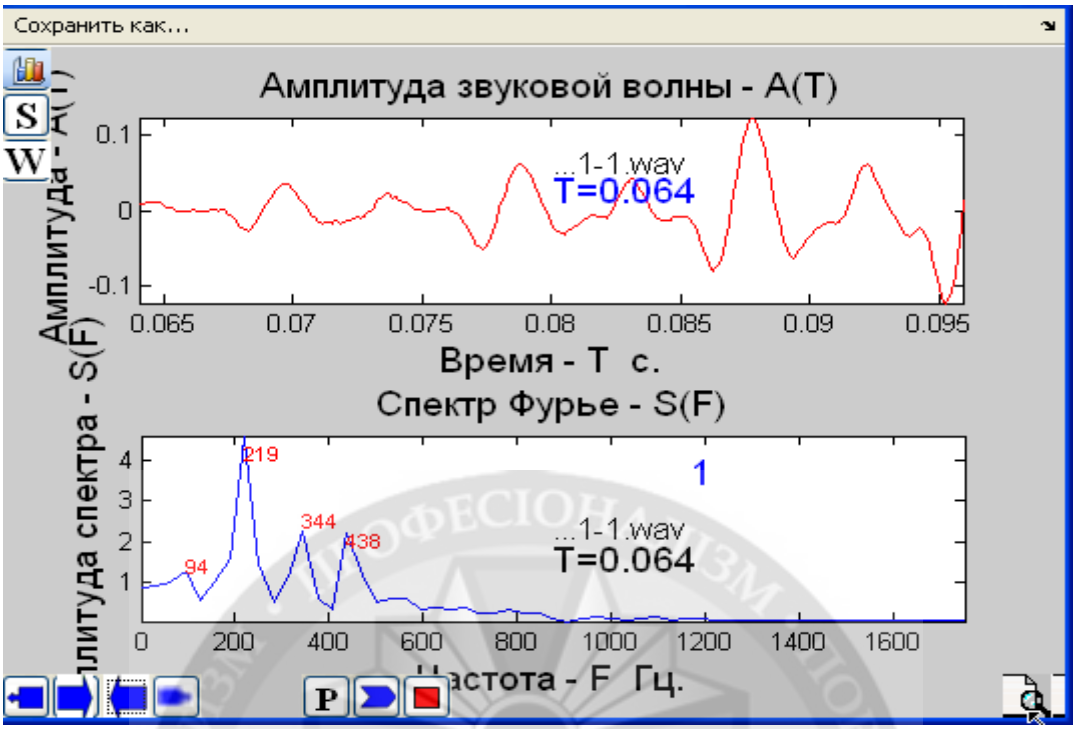

Рис.1. Спектр Фурье (ортогональное преобразование)

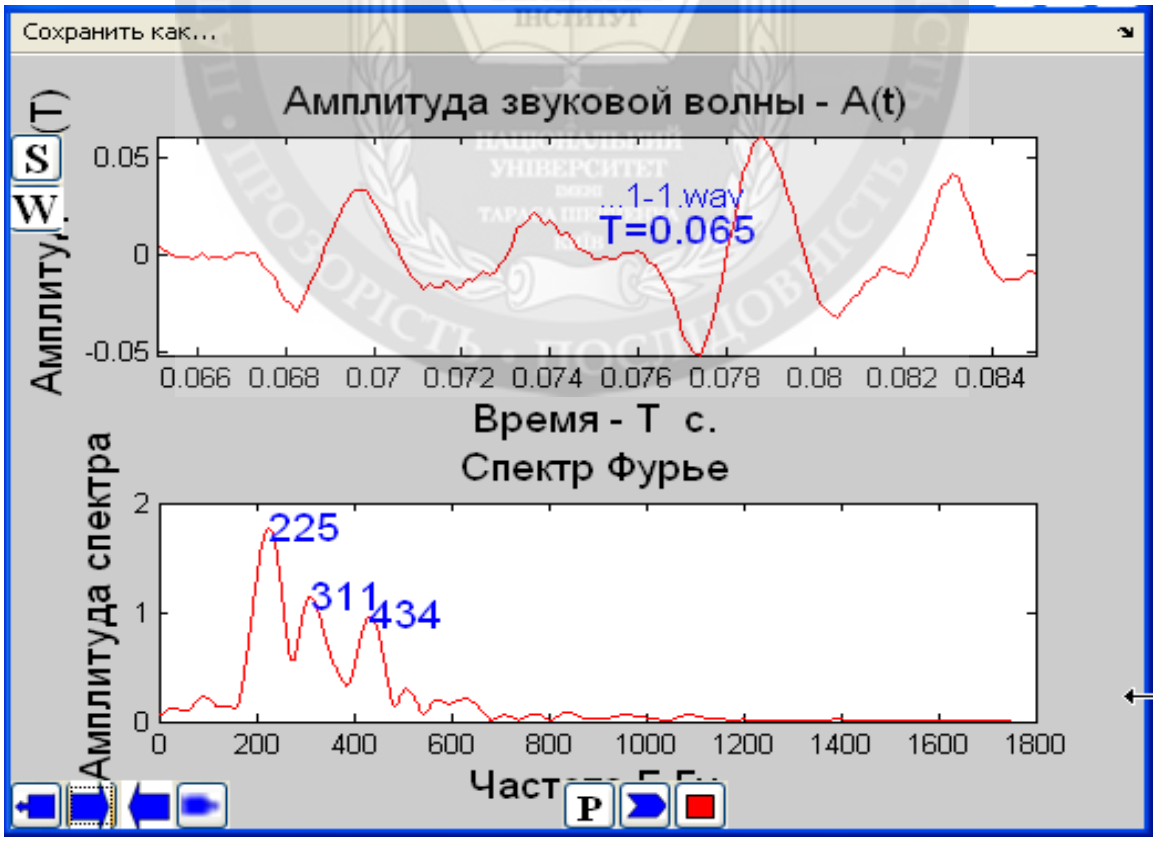

Рис. 2. Спектр Фурье (неортогональное преобразование)

В частности, на протяжении длительности фонемы локализация первых по величине максимумов амплитуды спектра изменяется в пределах нескольких герц по частоте (при шаге по частоте 1Гц). Исследования, результаты которых приведены в данной работе, проводились при соотношении сигнал шум не менее 20 дБ.

Более высокая стабильность спектральных характеристик для неортогонального базиса позволяет производить более устойчивые оценки параметров характеристик голоса. Так, например, на рис. 3 приведены графики распределения частоты основного тона речи (ЧОТ). Эти графики рассчитывались по среднему расстоянию по частоте между первыми 8 максимумами спектра для ортогонального и неортогонального вариантов для одного и того же файла с речью одного диктора.

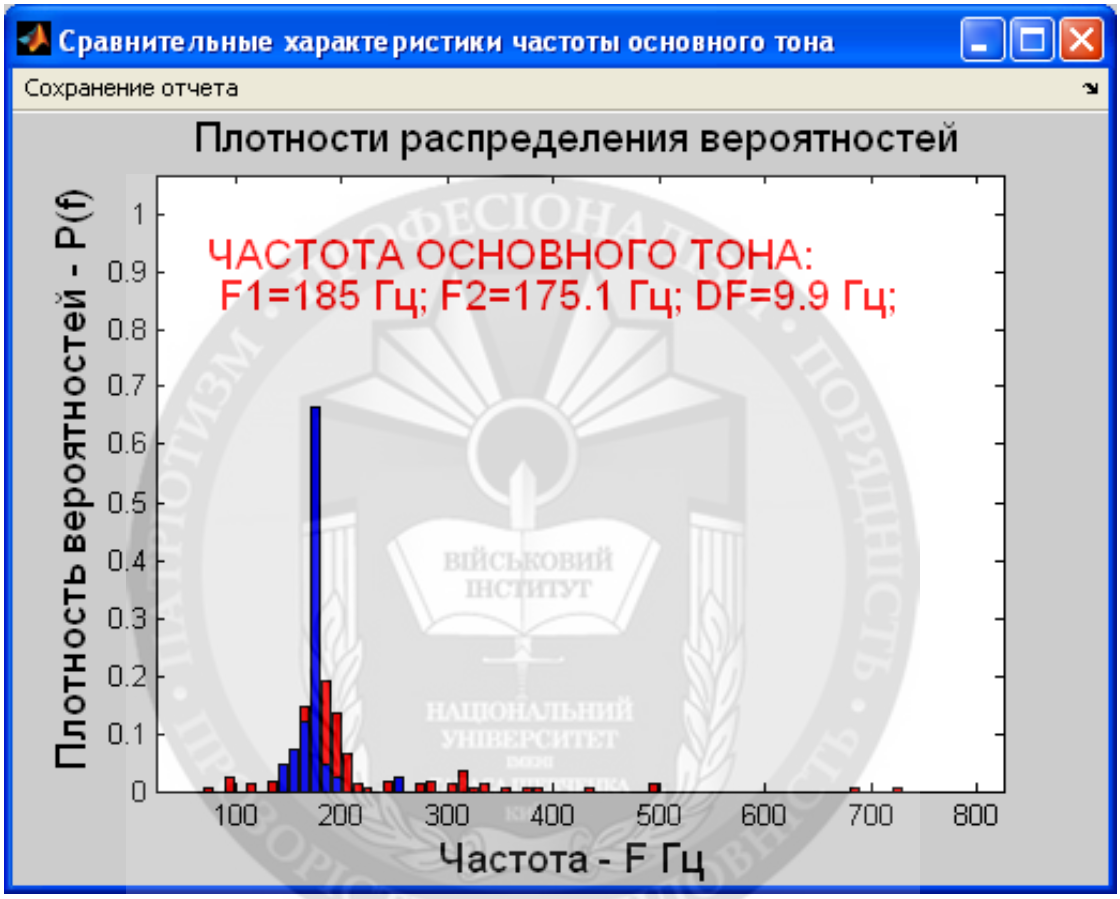

Рис. 3. Сравнительные распределения частоты основного тона

Из рис. 3 видно, что различие между распределениями, вычисленными на основе одного и того же алгоритма, весьма существенно. Для ортогонального варианта (красный график) дисперсия плотности распределения вероятностей гораздо выше. Это приводит к существенной разнице (DF=9,9 Гц) в оценке величины среднего значения ЧОТ F_1 и F_2 для одного и того же диктора.

Важным фактором неортогонального преобразования является существенно более высокая гладкость спектра в районе первых по величине локальных максимумов. Это позволяет реализовывать сравнительно простые эффективные алгоритмы расчета максимумов спектра. Кроме того, в ряде случаев разница в положении максимумов спектра для двух вариантов преобразований может составлять до 50 Гц.

Необходимо отметить следующее. Спектральное дискретное преобразование Фурье с ортогональным базисом вместе с обратным преобразованием обеспечивает взаимно однозначное соответствие между частотной и временной областями. Вариант неортогонального базиса является избыточным с точки зрения обратного преобразования.

Число временных отсчетов намного меньше числа отсчетов, которые можно получить в частотной области.

Известно, что при малых временных интервалах оконного преобразования для достаточно качественного прослушивания речи достаточно порядка 8 первых формант речи [2–7]. Аналогичную аппроксимацию амплитуды звуковой волны возможно реализовать на основе нескольких локальных максимумов для неортогонального базиса (1). Эксперименты показывают, что в этом варианте частотных характеристик для того же качества при прослушивании речи достаточно 4 – 5 формант. (При этом трактовка формант аналогична классическому варианту).

Вторым важным фактором стабильности спектральных характеристик для неортогонального базиса на малых временных интервалах является возможность построения независимых от характеристик голоса функций, зависящих от локальных максимумов спектра, которые могут являться устойчивыми идентификаторами звуков речи. Результаты этих исследований будут изложены в последующих статьях.

Выводы. Экспериментальные исследования речевых аудио файлов на основе преобразования Фурье с неортогональным базисом для временных интервалов порядка 20 мс показали высокую стабильность спектральных характеристик. Эта стабильность позволяет производить оценку важных речевых параметров речи в частотной области с более высокой степенью достоверности.

ЛИТЕРАТУРА:

1. Helmholtz H. von, Die Lehe von Tonempfindungen. Brannschweig, Vieweg, 1863.

2. Фланаган Дж. Анализ, синтез и восприятие речи: Пер. с англ./ Под ред. А.А. Пирогова. – М.: Связь, 1968. – 396 с.

3. Фант Гуннар. Анализ и синтез речи. Пер. с англ. В.С. Лозовского и Н.В.

4. Психоакустические аспекты восприятия речи. Механизмы деятельности мозга / Под. ред. Н.П. Бехтеревой. — М.: Наука, 1988. -504 с.

5. F.Elinek, 1976. Распознавание непрерывной речи статистическими методами. ТИИЭР 64, №4, с.131-160.

6. F.Elinek, 1985. Разработка экспериментального устройства, распознающего раздельно произнесенные слова. ТИИЭР 73, №11, с.91-99.

7. Цвикер Э., Фельдкеллер Р. Ухо как приемник информации. /Пер. с нем. под ред. Б.Г. Белкина – М.: Связь, 1971. – 225 с.

8. Алдошина И.А. Основы психоакустики. Звукорежиссер – 2000. – №6. – С. 36–40.

9. Сорокин В.Н. Теория речеобразования. – М. – Радио и связь. 1985. – 312 с.

10. Малла С. Вейвлеты в обработке сигналов / С. Малла. – М.: Мир, 2005. – 670 с.

11. Рыбальский О. В. В.И. Соловьев Система идентификации аппаратуры аудиозаписи на основе мультифрактального подхода. Вісник Східноукраїнського національного університету, № 9 (151), 2010. – С. 58–64.

12. Mandelbrot B. Statistical Methodology for Non-Periodic Cycles:From the Covariance to R/S Analysis. Annals of Economic Social Measurement 1, 1972.

13. Mandelbrot B. The Fractal Geometry of Nature. New York: W. H. Freeman, 1982.

14. Mandelbrot B. A Multifractal Walk Down Wall Street. Scientific American, 1999.

15. Mandelbrot B.B. Robustness of the rescaled range R/S in the measurement of non-cycling long-run statistical dependence // Water Resources Research. 1969. V. № 5. P. 967-988.

16. Павлов А. Н. Мультифрактальный анализ сложных сигналов / А. Н. Павлов, В. С. Анищенко // Успехи физических наук. – 2007, Том 177, №8.

17. Федер Е. Фракталы / Е. Федер. – М.: Мир, 1991. – 326с.

Рецензент: д.т.н., проф. Лєнков С.В., начальник науково-дослідного центру Військового інституту Київського національного університету імені Тараса Шевченка

к.т.н., доц. Соловйов В.І., д.т.н., проф. Рибальський О.В.

СПЕКТРАЛЬНИЙ АНАЛІЗ І МОВНІ ТЕХНОЛОГІЇ

У статті розглянуті результати досліджень авторів особливостей спектрального аналізу сигналів мови на основі ортогонального і неортагонального часо- частотного перетворення. Подібний підхід спирається на результати сучасних досліджень в області нейрофізіології і психофізіології слуху. Приведені результати досліджень по застосуванню часо-частотного перетворення на основі неортогонального базису для аналізу мовних файлів. Показано, що спектри на основі такого перетворення дозволяють отримувати достовірніші оцінки параметрів мови в частотній області. Отримані результати можуть бути застосовані при побудові систем ідентифікації диктора за параметрами його голосу та при побудові систем автоматичного розпізнавання мови. Крім того, отримані результати можуть зробити істотний вплив на побудову систем захисту мовної інформації з обмеженим доступом.

Ключові слова: частотний спектр, ортогональні базис, неортогональний базис, фонема, форманта, характеристики голосу.

V. Solovyov, O.Rybalsky

SPECTRAL ANALYSIS AND SPEECH TECHNOLOGY

The results of researches of authors of features of spectrology of signals of speech on the basis of ortogonal and neortogo-regional time of frequency transformation are considered in the article. Similar approach leans against the results of modern researches in area of neyrophysiology and physiopsychology of rumor. Results over of researches are brought on application of transformation from temporal in the frequency form of presentation of signals on the basis of unortogonal base for the analysis of speech files. It is shown that spectrums on the basis of such transformation allow to get more reliable estimations of parameters of speech in a frequency area. The got results can be applied at the construction of the systems of authentication of announcer on the parameters of his voice and at the construction of the systems of AVR. In addition, the got results can render substantial influence on the construction of the systems of defence of speech information with the limited access.

Keywords: frequency spectrum, a base, unortogonal base, phoneme, formant, descriptions of voice, is ortogonal.