КОМП'ЮТЕРНІ ЗАСОБИ, МЕРЕЖІ ТА СИСТЕМИ

E.N. Chichirin

SUBSYSTEM PROCESSING AND FORMING AUDIO SIGNAL

It was designed simulation environment for real-time processes frequency and time processing audio signals. The expediency of development time and neural network of speech processing techniques are considered.
Key words: audio system, speech processing, neural networks.

Розроблено середовище для моделювання в реальному часі процесів частотної і часової обробки аудіо сигналів. Обгрунтована доцільність розвитку часових і нейромережевих методів обробки мови.
Ключові слова: аудіо системи, обробка мови, нейромережі.

Разработана среда для моделирования в реальном времени процессов частотной и временной обработки аудио сигналов. Обоснована целесообразность развития временных и нейросетевых методов обработки речи.
Ключевые слова: аудио системы, обработка речи, нейросети.

УДК 681.3(031)

Е.Н. ЧИЧИРИН

ПОДСИСТЕМА ОБРАБОТКИ И ФОРМИРОВАНИЯ АУДИО СИГНАЛОВ

Методы и средства обработки и формирование аудио сигналов всегда представляли научный и практический интерес как в чисто информационных, так и в различных производственных (речевое управление, эхо- и гидролокация [1, 2], техническая диагностика, акустический дизайн и т. п.) технологиях.

Применительно к основной, связанной с рассматриваемыми вопросами областью, а именно, обработкой и формированием речевых последовательностей сигналов, следует отметить, что мы присутствуем при очередной смене парадигмы синтеза речи, представленной недавно приобретенной Google компанией DeepMind.

Развиваемую в течении нескольких десятилетий последовательность речевых технологий, как то фонемный и формантный синтез, мел-кепстральное представление и ЛПК-кодирование, а также чисто компилятивные методы "сшивания" речи из библиотек ее естественных кусочков DeepMind дополнила алгоритмами "попиксельного" формирования речи рекурентной сверточной нейросетью WaveNet [3]. Обучение сети проводилось на текстах, озвученных носителями языка.

За счет учета временного и нестационарного частотного контекстов, представленных каждым из цифровых отсчетов на протяжении 2–3 фонем и лингвистических особенностей текста на уровне слов и предложений удалось сократить разрыв в баллах между искусственной и естественной речью на 50 % по сравнению с существующими параметрическими и компилятивными методам. .

В конечном итоге, стремление учитывать эмоциональную окраску и нестационарный характер речи при любом сочетании диалоговых пар человек - (ро)бот приведет к необходимости и при ее распознавании применять глубокие нейросетевые методы.

С другой стороны, размерность совокупности матриц весовых коэффициентов современных нейросетей помимо практической невозможности их адекватного логического восприятия требуют для их обработки огромных вычислительных мощностей, особенно на стадиях обучения. Решением может стать сохранение отработанных методов параметрического моделирования, например, на квазистационарных участках речевого сигнала.

В любом случае моделирование и отработка совокупности процессов анализа и синтеза речевых сигналов в масштабе времени реального диалога наряду с возможностью дополнения их, например, элементами поэтапного ("естественного") нейросетевого обучения является весьма актуальным.

Рассматриваемая далее программная реализация аудио подсистемы содержит средства первичной обработки и формирования речевых сигналов, имеет встроенную поддержку дуплексного речевого взаимодействия и предназначена для поэтапного решения вышеизложенных задач.

Структурная схема подсистемы обработки аудио сигналов (ПОАС) показана на рис. 1. ПОАС содержит:

- панель подготовки и исполнения (макро) программ;
- библиотека алгоритмов обработки и формирования аудио сигналов;
- панель управления банком операционных регистров;
- графический монитор состояний операционных регистров;
- нейросетевая база данных – NNetDBase;
- панель управления NNetDBase;
- блок обучения и нейросетевой обработки;
- панель установки параметров обработки и ввода-вывода – PSet;
- контроллер процессов обработки и ввода-вывода;
- диагностический блокнот – NoteBook.

ПОАС разрабатывалась как средство экспериментального моделирования с различными существующими и новыми алгоритмами звуковой обработки. В силу большого их числа и разнообразия элементы управления ими (выделение ресурсов памяти, времени и экранного места, установка режимов и параметров исполнения, элементы пуска-останова) по возможности унифицированы и сведены в несколько панелей. Более крупные функциональные формы отмечены в пунктах главного меню на рис. 2.

Архитектура и интерфейс ПОАС поддерживают возможность гибкого использования внутрисистемных ресурсов, представленных банком операционных регистров и средствами графического мониторинга их состояния, виртуальной страничной памятью нейросетевой базы данных NNetDBase, а также расширяемой библиотекой алгоритмов обработки и формирования аудио сигналов.

Панель управления банком регистров обеспечивает фиксированное, ручное и автоматическое назначение регистров отдельным приложениям из библиотеки алгоритмов, подключение их к средствам оперативной визуализации и отображение дополнительной технологической информации. Банк регистров поддерживает динамическое распределение оперативной системной памяти и ее очистку совместно с сопутствующей информацией в панели управления и графическом мониторе по завершению работы приложений.

Нейросетевая база данных (NNetDBase) использует механизм динамической подкачки страниц виртуальной памяти для отображения дисковых файлов в оперативном адресном пространстве исполняемой подпрограммы. В ПОАС планируется поддержка произвольных нейросетевых структур, с перспективой возможности их реконфигурации в процессе обучения.

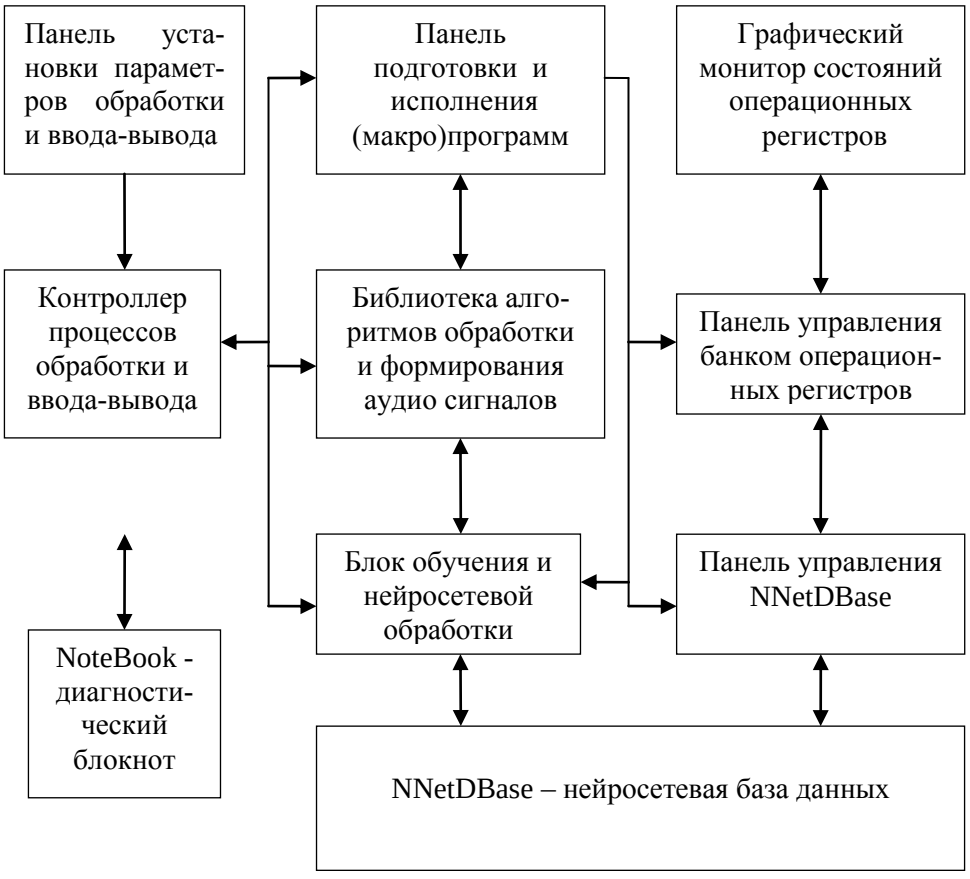

РИС. 1. Структурная схема подсистемы обработки аудио сигналов

Обновляемая библиотека алгоритмов обработки и формирования аудио сигналов содержит подпрограммы временного и частотного анализа моно- и стерео последовательностей, а также обратного преобразования полученных отображений в звуковой сигнал. В частности, библиотека включает в себя алгоритмы прямого и обратного преобразования Фурье (FFT – Fast Fourier Transform), ортогонального вейвлет преобразования Дебеши, а также алгоритмы прямой интерполяции звуковых последовательностей во временной области. Источник сигнала – микрофон или wav-файл, приемник – звуковой излучатель или wav-файл. Поддерживается непрерывный дуплексный режим работы. Варианты нейросетевого обучения и обработки в стадии разработки.

Элементы пуска-останова подпрограмм показаны на рис. 2 мнемоническими их наименованиями в ячейках таблицы имеющихся в библиотеке алгоритмов. Подпрограммы не имеющие конфликтов фиксировано назначенных одноименных регистровых ресурсов можно объединять в (макро)программы путем автоматического сохранения последовательности запускаемых в ручном режиме подпрограмм.

При надлежащем оформлении программы допускается пошаговое их исполнение в диагностическом режиме с сохранением производительности в непрерывном режиме выполнения.

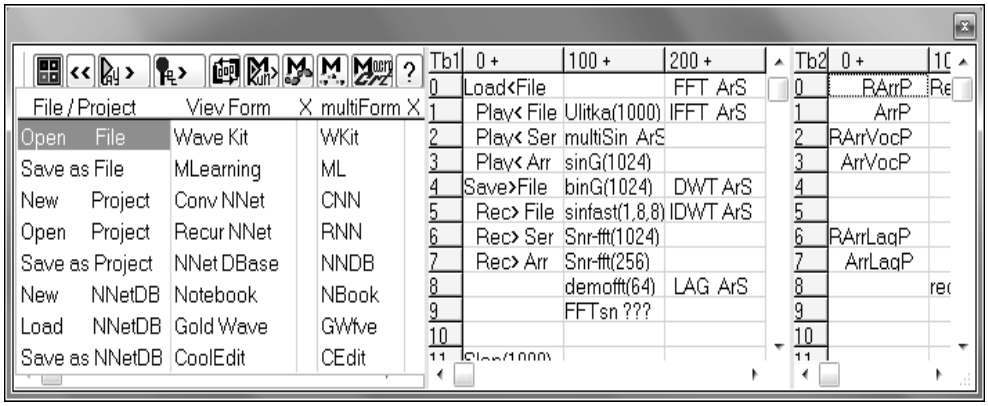

РИС. 2. Главное меню и таблица подпрограмм на панели управления Wave-конструктора

Контроллер процессов обработки и ввода-вывода обеспечивает взаимную синхронизацию входных и выходных последовательностей сигналов с их низкоуровневой и высокоуровневой обработкой в реальном масштабе времени речевого обмена.

Панель установки параметров обработки и ввода-вывода PSet содержит элементы настройки системных ресурсов (длительность, частота дискретизации) и режимов выполнения прикладных программ (параметры буферизации, визуализации, варианты функционирования, оконного взвешивания и т. п.).

Диагностический блокнот NoteBook содержит средства фиксации дополнительной рабочей и отладочной информации в графической и табличной форме.

В совокупности перечисленные средства составляют содержимое и определяют возможности волнового конструктора WaveKit ПОАС, предназначенного для моделирования процессов обработки звуковых, в том числе речевых, последовательностей. Интерфейс WaveKit с графическим монитором, панелью выполнения программ и диагностическим блокнотом показан на рис. 3.

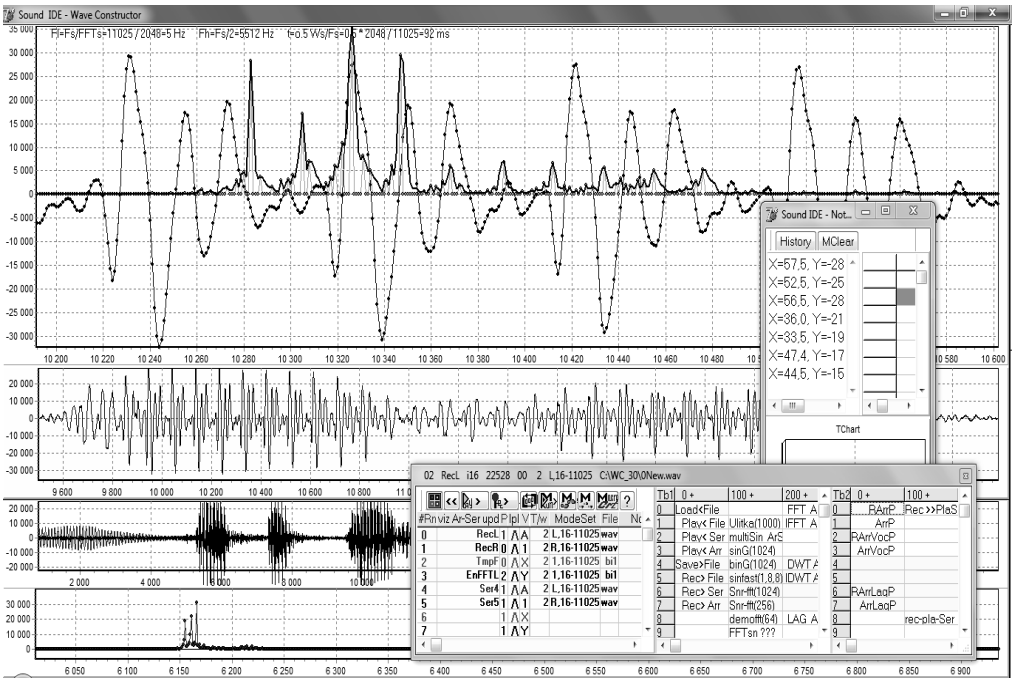

РИС. 3. Окно WaveKit для обработки и формирования аудио сигналов

Показанная на рис. 4 архитектура WaveKit изначально обусловлена функциональными потребностями визуализации и диагностирования процессов обработки лингвистически связанных участков речи, например, предложения. Исходя из выбираемых в PSet длительностей (в сек.) $Trec$ и $Tplay$ ближнего лингвистического контекста для входных и выходных участков речи определяются минимально допустимые величины основных структурных параметров – размеров (в цифровых отсчетах и байтах) $SCrec$ и $SCplay$ конвейерных регистров RCrec и RCplay для циклической буферизации процессов оперативной обработки принимаемой и формируемой последовательностей отсчетов

$$SCrec = ((Trec \cdot Fdrec + SBrec - 1)/SBrec) \cdot SBrec; \qquad (1)$$

$$SBplay = ((Tplay \cdot Fdplay + SBplay - 1) / SBplay) \cdot SBplay. \qquad (2)$$

Здесь "/" – операция целочисленного деления, $SBrec$, $SBplay$ – размеры буферных регистров ввода-вывода RBrec, RBplay устанавливаемых в PSet на минимально допустимых для устойчивой работы значениях, $Fdrec$, $Fdplay$ – частоты дискретизации ввода и вывода.

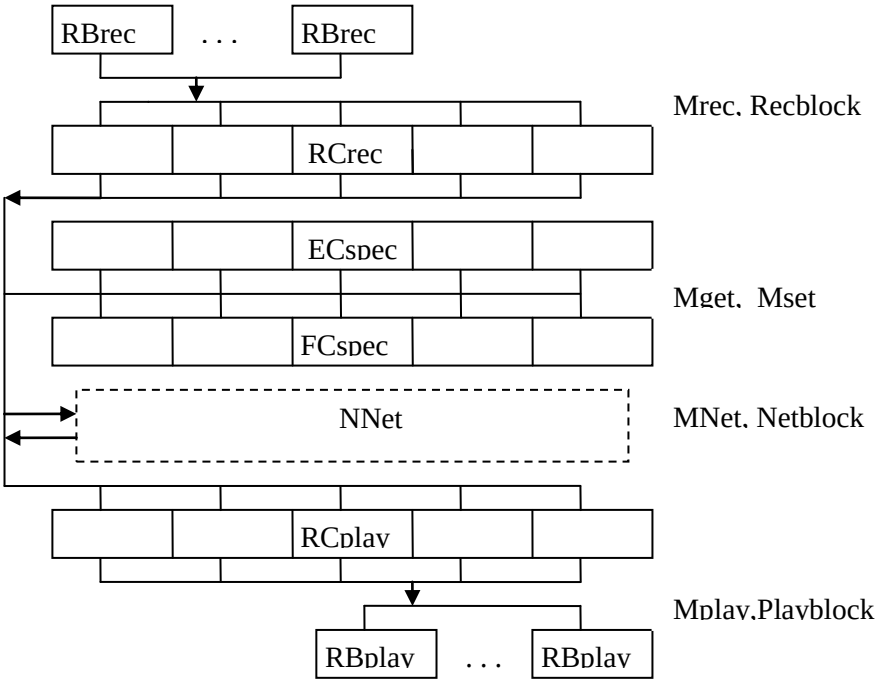

РИС. 4. Регистровая архитектура конструктора WaveKit

Промежуточные в конвейерной цепи регистры иной функциональности и формата представления данных (спектр – REspec и RFspec, фонетическая или иная разметка, текст) имеют согласованный с $SCrec$ и $SCplay$ размер для возможности визуального сопоставления их графических представлений.

Синхронизация одновременно исполняемых потоков кольцевой загрузки и обработки блоков каждого из конвейерных регистров осуществляется с помощью системных прерываний ввода-вывода, пользовательских прерываний, циклических индексов Mrec, Mset, Mget, MNet, Mplay и маркеров готовности Recblock, Netblock, Playblock очередного блока. Дополнительная функциональность обеспечивает "бесшовную" стыковку соседних блоков при необходимости их совместной во времени обработки.

Частоты дискретизации $Fdrec$, $Fdplay$ (далее Fd) в (1, 2) устанавливаются в PSet исходя из чувствительности слуха в частотной и временной областях и

особенностей FFT-преобразования. При частоте дискретизации Fd и кратной степени 2 размерности N FFT (числу его входных отсчетов) максимальная $f\max$ и минимальная $f\min$ выявляемые во входной выборке спектральные частоты определяются соотношениями

$$f\max = Fd/2 \text{ Гц};$$

$$f\min = \Delta f = Fd/N = 2f\max/N \text{ Гц};$$

$$\Delta t = T\min = 1/f\min = \cdot N/Fd \text{ сек.} \quad (3)$$

Здесь $\Delta f = f\min$ – разрешение (шаг) по частоте спектра. Энергетический спектр содержит ряд действительных (без учета зеркальных) значений модулей возрастающих с шагом Δf частотных составляющих входной последовательности, усредненных во времени в пределах длительности Δt выборки ее отсчетов. То есть длительность Δt выборки – это разрешение (шаг) по времени спектра. Расширение частотного диапазона $[f\min,..., f\max]$ и улучшение разрешение Δf по частоте спектра за счет изменения Fd и N приводит к ухудшению разрешения Δt во времени, как приведено в таблице. Для доказательства имеющего место принципа неопределенности достаточно переписать соотношение (3) в виде $\Delta t \cdot f\min = \Delta t \cdot \Delta f = 1$.

Временное разрешение слуха около 20 мсек., и типично выбираемые в панели установки параметры оцифровки для речи и музыки отмечены в таблице курсивом.

ТАБЛИЦА

N Fd	64	128	256	512	1024	2048	4096	8192	Δf = $f\min$
8000	125 4000 8	62,5 4000 16	31,2 4000 32	15,6 4000 64	7,8 4000 128	3,9 4000 256	1,95 4000 512	0,98 4000 1024	fmin fmax Δt мсек
11025	172,3 5012 5,8	86,2 5012 11,6	43,1 5012 23,2	*21,5* *5012* *46,5*	10,76 5012 92,9	5,38 5012 185,8	2,5 5012 371,6	1,25 5012 743,2	fmin fmax Δt ьсек
22050	344,6 11025 2,9	172,3 11025 5,8	86,2 11025 11,6	43,1 11025 23,2	21,5 11025 46,5	10,76 11025 92,9	5,38 11025 185,6	2,69 11025 371,2	fmin fmax Δt мсек
44100	689,1 22050 1,45	344,6 22050 2,9	172,3 22050 5,8	86,2 22050 11,6	43,1 22050 23,2	*21,5* *22050* *46,5*	10,76 22050 92,9	5,38 22050 185,6	fmin fmax Δt мсек

За счет оконного взвешивания с перекрытием 1/2 величина Δt снижается до 23,3 мсек. Из доступных шести окон лучшие результаты у окна Хемминга.

Суммарная латентность, обусловленная размерами блоков ввода-вывода и других уровней низкоуровневой конвейерной обработки составляет 0,2 сек.

Преобразование входного сигнала в частотный спектр в ПОАС представлено двумя FFT-алгоритмами, первый из которых не требует памяти для хранения вращающих коэффициентов, а второй обеспечивает ускоренную обработку действительных сигналов. Кроме того для улучшения временного разрешения Δt он поддерживает работу с дополненной нулями выборкой за пределами окна взвешивания. Как и ожидалось, это сопровождается размытием низкочастотных спектральных линий.

На текущем этапе не преследовалась цель сжатия (вплоть до распознавания) речи. Исследовались возможности построения более производительных подходов ее восстановления в реальном времени с допустимым уровнем достоверности. Для спектрального внутреннего представления это оказался метод компиляции выходной последовательности суммой основных таблично генерируемых гармоник с сохранением фазы и упорядоченной близости частот конкатенируемых пар на стыках соседних блоков. При сопоставимых затратах памяти асимптотически линейный (в отличии от FFT) метод временного кодирования экстремумов входного сигнала и его простого интерполирования 1/4 периодами соответствующих гармоник показал тем не менее лучшее качество восстановления. Очевидно, это происходит за счет эффекта низкочастотного маскирования высокочастотных составляющих речи и шума, особенно при высоком пороге отбора экстремальных значений.

Выводы. Разработана подсистема для экспериментального моделирования процессов обработки и формирования аудио сигналов в реальном масштабе времени их поступления. Проведены исследования различных методов обработки речи, основанных на амплитудно-частотном и амплитудно-временном представлении и восстановлении речевых последовательностей. Получены подтверждения практической эффективности и целесообразности дальнейшего развития временных подходов, как наиболее перспективных в плане поддержки их нейросетевыми методами обучения в реально протекающих диалоговых процессах.

1. *Палагин А.В., Семотюк М.В., Чичирин Е.Н., Сосненко К.П. .А.* Acoustic commander – интегрированная операционная среда для измерения и расчета акустических параметров. *Комп`ютерні засоби, мережі та системи*. К.: Ін-т кібернетики імені В.М. Глушкова НАН України. 2009. № 4. С. 3–10.
2. *Винцюк Т.К.* Анализ, распознавание и интерпретация речевых сигналов. К.: Наукова думка, 1987. 262 с.
3. *WaveNet* - Generative model for Raw Audio. *Available at https:*//deepmind.com/blog/wavenet-generative-model-raw-audio.

Получено 19.10.2016