УДК 681.324

О.И. Данченков, Д.В. Николаенко
Государственный университет информатики и искусственного интеллекта, г. Донецк, Украина
Автомобильно-дорожный институт государственного высшего учебного заведения
«Донецкий национальный технический университет», г. Донецк, Украина

Использование динамических портретов звука при распознавании речевого сигнала

Рассматривается научная проблема распознавания речевых образов. Проведен анализ современных компьютерных средств голосового управления. Сформулирована структура пользовательского звукового интерфейса. Предложено использование динамических портретов звука как части процесса определения параметров анализируемого звукового сигнала.

Общая постановка проблемы

Дальнейшее распространение использования различных информационных систем приводит к необходимости предоставления пользователю максимальных удобств при работе с компьютером в режиме диалога. Тенденции совершенствования коммуникационного интерфейса ведут к упрощению диалога пользователя с ЭВМ. В последние годы разработке удобного интерфейса уделяется пристальное внимание со стороны ведущих производителей программных продуктов. Привычным стандартом стали многооконные системы, оснащенные визуальными средствами управления в соответствии с принципами GUI (Graphical Users Interface). Управление информационными системами больше не требует поиска нужной клавиши на клавиатуре. Все осуществляется наглядно, и пользователь видит результаты своих действий на мониторе компьютера, в любой момент он может обратиться к системе помощи, которая стала неотъемлемым компонентом любой информационной структуры.

В конечном итоге интерфейс пользователя компьютерной системы должен обеспечивать возможность общения с ней на естественном языке, в том числе и с помощью речи. На рис. 1 приведена возможная структура вычислительной системы со звуковым интерфейсом.

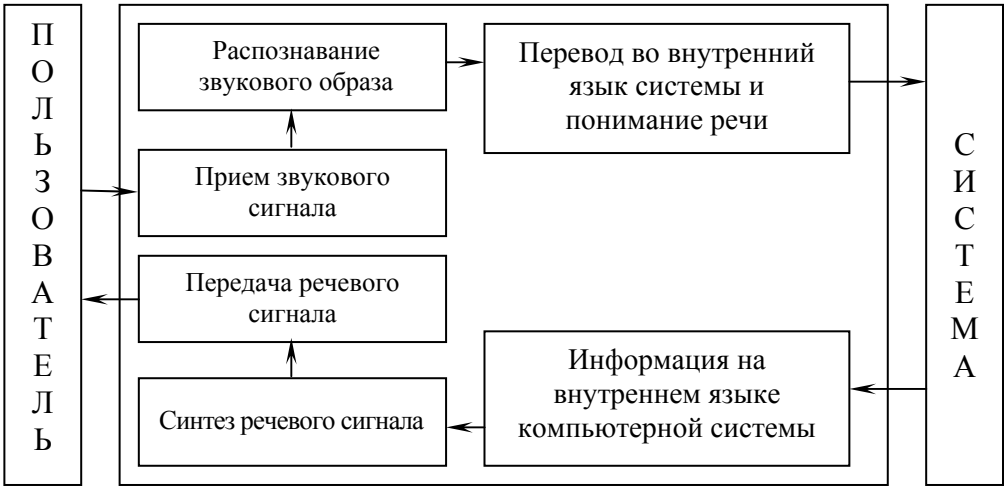

Рисунок 1 – Структура звукового интерфейса

В настоящее время следует отметить преимущественные успехи в решении задач синтеза звука по сравнению с распознаванием звуковых образов и понимания речи.

Тем не менее, уже сейчас можно выделить ряд областей, где применяется анализ звука и речи. Упомянем лишь некоторые из них, где производится измерение параметров речи: биометрия, судебная экспертиза, медицина. Голос человека можно использовать как пропуск в системах с ограничением доступа (например, в автоматическом контрольно-пропускном пункте, КПП). При производстве судебной экспертизы материалов звукозаписи часто нужно провести идентификацию личности. Можно определять эмоциональное состояние (уровень стресса) по параметрам устной речи. Такой способ имеет то преимущество, что к человеку не нужно присоединять датчики.

Анализ современных средств речевого управления

Существует ряд программных продуктов, позволяющих пользователю осуществлять ввод текста голосом; управлять голосом периферийным оборудованием; осуществлять голосовое управление отдельными функциями операционных систем; осуществлять голосовое управление функциями текстовых редакторов и прикладных программ; оформлять документы (включая формулы). Среди таких программ можно выделить:

– Aria Listener фирмы Prometheus products;
– «Горыныч» фирмы VoiceLock;
– IN3 Pro Voice Command корпорации Command Corp;
– Listen компании Verbex Voice Systems;
– QuickSwitch фирмы BitWare Consulting.

Также имеется ряд программ для диктовки – речевой ввод текстовой и цифровой информации, например:

– NaturallySpeaking Preferred (Dragon Systems);
– Via Voice 98 Executive Edition (IBM);
– Voice Xpress Professional (Lernout & Hauspie);
– FreeSpeech 98 (Philips).

К основным недостаткам этих программ можно отнести следующие [1, с 5]:

– диктовка должна осуществляться по словам, то есть после каждого слова нужно делать паузу, что не совсем удобно и понижает скорость набора текста;
– длительная настройка программы интерфейса на конкретного диктора, обучение системы, для получения некоторой базы слов (иногда достаточно большой). Например, для программы «Горыныч» фирмы VoiceLock этот объем составляет 5000 слов, а в коммерческой версии – 10000, причем эта база может постоянно пополняться;
– достаточно высокая цена.

Существует множество подходов и методов решения задачи распознавания речи. Выделим из них три основных метода [2-4]:

– использование искусственных нейронных сетей как мощного средства распознавания образов;
– использование спектрального представления сигнала для выделения фонем звуков в слитной речи;
– метод линейного предсказания.

Актуальным является вопрос первичного описания речи, поиска таких форм его представления, которые обеспечивали бы простое и надежное выделение информативных признаков сигнала.

Для эффективного акустического анализа важно, с каким представлением исходного сигнала работает система автоматического распознавания речи, какие параметры выделяются для последующей фонетической обработки и как эти параметры могут быть надежно выделены в речевом сигнале.

Основной идеей настоящей работы является использование динамических портретов звука как составной части процесса автоматического распознавания речи и позволяющих решить научно-техническую задачу определения основных параметров анализируемого звукового сигнала.

Построение динамических портретов звука

Динамические портреты звукового сигнала – достаточно удачная форма представления речи, позволяющая выявить наиболее характерные, сравнительно инвариантные свойства звуков, различительные признаки отдельных звуков и их классов [4].

Динамический портрет звука состоит из трех составляющих:

– массив максимальных значений (контур интенсивности) – все значения отчетов (выборок) параметрического кода сигнала заменяются значением максимального отсчета на интервале времени Т определенной длительности. Полученный массив нормируется по амплитуде для всего анализируемого отрезка речевого сигнала;

– контур числа переходов через ноль – подсчитывается число переходов через ноль на том же интервале времени Т, что и в случае контура интенсивности;

– контур числа нулей – подсчитывается число нулей на интервале времени Т.

Алгоритм распознавания звукового образа с использованием динамических портретов звука можно представить в виде последовательности следующих этапов:

1. Аналоговый сигнал из микрофона попадает на вход звуковой карты.
2. В звуковой карте аналоговый сигнал преобразуется в цифровой. При этом программа звукозаписи при помощи драйвера звуковой карты генерирует звуковой файл формата WAV.
3. Данные из этого файла с помощью специальной программы обрабатываются и на основании этих данных строится динамический портрет, который может быть выведен на экран.
4. Дальнейшая обработка состоит в анализе динамических портретов звука с целью выделения образов фонем по специальному алгоритму.
5. По выделенным фонемам может быть восстановлен текст, который диктовался человеком.

Если первые две задачи решаются стандартными программными средствами, то для разработки алгоритма в задаче 4 необходимо создание специального АРМа, который бы позволял исследовать динамические портреты звука.

Такая программа должна позволять на базе современных технических средств и стандартов представления акустической информации в персональном компьютере создавать динамические портреты из любого оцифрованного звука для последующего анализа речевых сигналов и выявления информативных признаков фонем звучащей речи и построения технических систем, использующих речевое управление.

В качестве технических средств исследования речевого сигнала может быть выбран мультимедийный набор персонального компьютера, в состав которого входит звуковая карта, позволяющая оцифровывать любой звук в диапазоне частот от 8000 Гц до 48000 Гц.

В результате проведенных экспериментальных исследований была выбрана частота дискретизации сигнала 44100 Гц, что объясняется психофизическим эффектом сглаживания в слухе [4]. 16-битная оцифровка в совокупности с частотой дискретизации, равной 44100 Гц, дает предельное соотношение «сигнал/шум» около 98 дБ.

Дискретизация сигнала позволяет осуществлять комплексное исследование речевого сигнала, в частности, решение задачи идентификации по голосу, предполагающее обработку тонкой временной структуры сигнала. Нижний порог частоты дискретизации определяется на основании теории В.А. Котельникова и не превышает 20 кГц [5].

Для хранения оцифрованного звука был выбран формат файла WAV, преимущества которого заключаются в отсутствии компрессии файла, что позволяет осуществлять прямой доступ к данным без предварительной декомпрессии, а также распространенности формата, надежности хранения данных, наличии наиболее полной технической документации. Структура формата WAV-файла приведена в табл. 1.

Таблица 1 – Структура WAV-файла

Смещение	Длина	Описание
0h	4h	Идентификатор формата ('RIFF')
4h	4h	Длина блока данных (длина файла – 8h)
8h	4h	Идентификатор блока звуковых данных ('WAVE')
0ch	4h	Идентификатор подблока заголовка ('fmt" – с пробелом в конце)
10h	4h	000ch/0010h – длина подблока заголовка
14h	2h	01h – тип формата представления данных
16h	2h	Число каналов (1 – моно, 2 – стерео)
18h	2h/4h	Частота дискретизации, Гц
1ah/1ch	2h/4h	Скорость передачи данных, байт/с (произведение числа каналов, частоты дискретизации и разрядности в байтах)
1ch/20h	2h	Число байт для представления одного отсчета (1 – 8 бит моно, 1 – 16 бит стерео)
1eh/22h	2h	Разрядность, бит (8, 16)
20h/24h	4h	Идентификатор подблока данных ('data')
24h/28h	4h	Длина звуковых данных
28h/2ch		Звуковые данные

В качестве языка программирования был выбран язык высокого уровня Delphi 6, что объясняется наличием в этом языке программирования всех необходимых библиотек и методов для объектов, используемых в программе. Для иссле-

дования была написана программа, позволяющая получать динамические портреты речевого сигнала. Полученные портреты совпадают с динамическими портретами, полученными в работе [4]. На рис. 2 представлена исходная форма звукового сигнала, а на рис. 3 приведен динамический портрет этого звукового сигнала. В качестве звукового сигнала было использовано слово «сочиться».

АРМ имеет окно с двумя закладками. Одна – для отображения динамического портрета, другая – для отображения исходной (реальной) волны речевого потока. АРМ позволяет сохранять полученный динамический портрет в формате BMP-файла, распечатывать на принтере, масштабировать изображение динамического портрета, отображать экстремумы либо точками, либо в виде числовых значений для удобства анализа и восприятия.

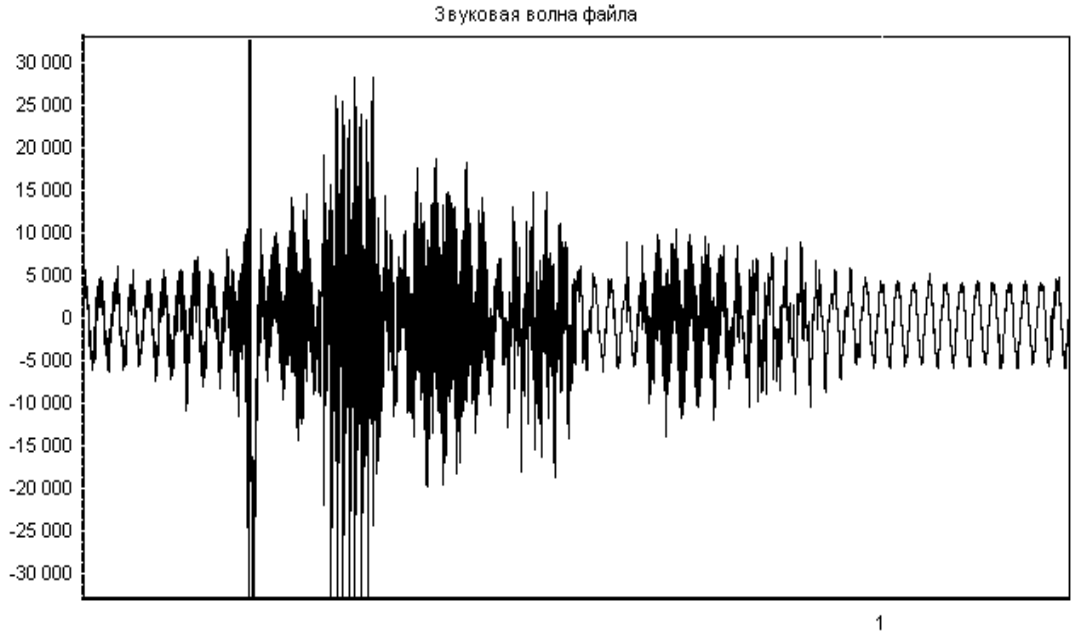

Рисунок 2 – Исходная форма слова «сочиться»

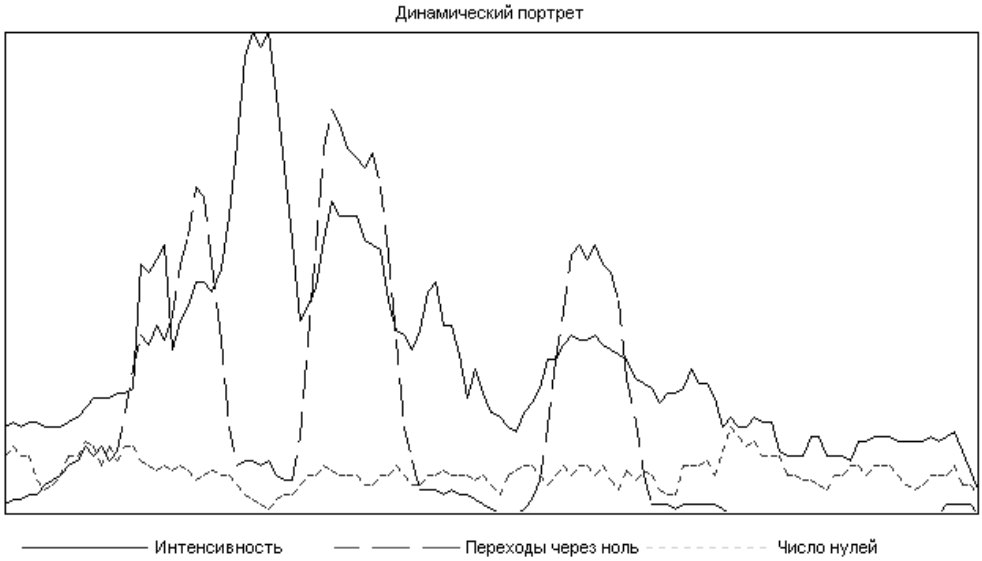

Рисунок 3 – Динамический портрет слова «сочиться»

Заключение

Проведенный анализ существующих методов распознавания речи показал, что одним из возможных методов распознавания звукового образа является алгоритм с использованием динамических портретов речевых сигналов. Полученные результаты построения динамических портретов позволили на практике определить параметры дискретизации речевого сигнала с учетом психофизического эффекта их сглаживания в слуховом аппарате человека. Результаты работы в перспективе могут быть использованы при решении научных проблем акустического анализа речи.

Литература

1. Режим доступу: www.art.bdk.com.ru/govor.
2. Федяев О.И., Гладунов С.А. Речевая компонента в интерфейсах информационных систем // Сб. научных трудов Донецкого национального технического университета. Серия «Информатика, кибернетика и вычислительная техника». – 2001. – С. 100-105.
3. Доросинский Л.Г., Николаев А.Н. Особенности применения методов распознавания речи в задачах анализа дефектов речи. – Режим доступа: http://www.sakrament.com/it-rus/index.htm.
4 Николаенко В.Л. Математические модели согласных сегментов речевого потока и их использование в системах автоматической обработки речи. – Харьков, 1988.
5 Технология распознавания голоса. – Режим доступа: http://www.cs.msiu.ru/projects/kurs/1999/9311/AI_CURSOVIK/kozlowa/docum2.html.

О.І. Данченков, Д.В. Ніколаєнко

Використання динамічних портретів звуку при розпізнаванні мовного сигналу

Розглядається наукова проблема розпізнавання мовних образів. Проведений аналіз сучасних комп'ютерних засобів голосового управління. Сформульована структура користувацького звукового інтерфейсу. Запропоновано використання динамічних портретів звуку як частини процесу визначення параметрів звукового сигналу, що аналізується.

O.I. Danchenkov, D.V. Nikolaenko

The using of dynamic sound portraits for speech signal understanding

The scientific problem of recognition of speech images is considered. The analysis of modern computer instruments of voice control is carried out. The structure of the user sound interface is formulated. The structure of dynamic sound portraits as a part of process of definition of parameters of an analyzed sound signal is offered.

Статья поступила в редакцию 03.12.2007.