Materials of e−conference "Speech synthesis and analysis".
Kazan State University, Russia
received September 16 , 1997.
print revised October 19, 1997.

ON ANALYSIS OF HEARING IMAGE OF SPEECH SIGNAL

Alexander I. Egorov, Vladimir V. Dubrowsky

Irkutsk Computing Centre of Siberian Branch
of the Russian Academy of Sciences
134,Lermontov Str, Irkutsk, 664033, Russia
E−mail: egorov@icc.ccsoan.irkutsk.su

ABSTRACT

The paper consist of introduction, 3 part and conclusion. At first part the methodology of speech recognition was considered. Second part devoted to experimental investigations of subconcious processes under perception of hearing image. In consequance of investigations the complex content of vowels was revealed. At third part of the paper one of the variants of alphabet of elementary hearing images of russian speech was proved by experiments. In the end of the paper an integral conclusion is drew.

1. Information, that a person use to form and to classify the elementary hearing image is placed at the part of the speech signal corresponding to syllable of the verbal speech.
2. Training system of recognizing arbitrary dictor's speech can be done by forming of template of elementary hearing images.
3. The process of speech recognition can be represented in the form of string of classification procedures of elementary speech image.

The results mentioned at the paper used by authors for constraction of imitative model of hearing analysis of russian speech.

KEY WORDS: speech signal, hearing analysis, subconcious process, elementary hearing image, alphabet, imitative model.

ОБ АНАЛИЗЕ СЛУХОВЫХ ОБРАЗОВ РЕЧЕВОГО СИГНАЛА

А.И.Егоров, В.В.Дубровский
ИрВЦ СО РАН, Иркутск, Россия
egorov@icc.ccsoan.irkutsk.su

Введение.

Одной из актуальных задач речевой информатики продолжает оставаться задача создания дикторонезависимых систем распознавания речи. Задача дополнительно усложняется необходимостью работы со словарями произвольного объема. Существуют различные точки зрения на причины медленного продвижения в решении этой задачи [1,2]. В сообщении дана оценка современного состояния проблемы автоматического распознавания слуховых образов (АРСО) и приведены новые результаты, которые позволяют сомневаться в правомерности использования традиционных подходов к разработке систем распознавания речевых сигналов (РС). Полученные результаты могут быть использованы при создании систем классификации слуховых образов.

О методологии распознавания слуховых образов речевого сигнала.

Известно [3], что ключевой для АРСО считается проблема первичного представления речевых сигналов . В настоящий момент по этому вопросу нет какой−либо установившейся идеологии. В результате исследователи РС и разработчики систем АРСО при выборе первичного описания РС маневрируют между несколькими недостаточно обоснованными теоретическими позициями. Мы полагаем, что выдвижение первичного анализа речи в качестве ключевой проблемы АРСО является следствием использования традиционной

методологии проведения речевых исследований. В основу этой методологии положена идея о возможности только математическими методами осуществить распознавание РС по его первичному описанию.

При этом предполагается, что из первичного описания может быть

выделен набор некоторых информативных признаков, в пространстве

которых возможно проведение распознавания РС. Таким образом фактически подразумевается «эквивалентность» между слуховым образом − результатом субъективного восприятия речевого сигнала в слуховой системе (СС) человека и самим РС. Очевидно, что использование представлений о названной эквивалентности требует серьезного обоснования. И поскольку такого обоснования пока не существует, задачу классификации слуховых образов (СО) нельзя отождествлять с задачей распознавания РС.

Отметим также, что прямым следствием попыток непосредственно

распознать речевой сигнал является использование при анализе

сигнала большого количества теоретических представлений, постановок и методов решения задач, носящих искусственный характер. В этом отношении наиболее показательно предположение о существовании в устной речи аллофонов. При помощи аллофонов пытаются как−то учесть известное разнообразие и изменчивость РС, соответствующих восприятию в СС одного и того же звукотипа. Действительно, реализации РС, соответствующие слуховому восприятию одного и того же звукотипа, зависят от темпа и интенсивности РС, от позиции сегмента РС в слове, места ударения, функционального и эмоционального состояния диктора и множества других факторов. В слуховой же системе человека эти РС уверенно относятся к одному СО. Таким образом, подмена классификации СО на распознавание РС,

производимое по его описаниям, при выборе которых практически полностью игнорируется определяющее участие головного мозга в порождении и восприятии РС, по—видимому, не может привести к раз—работке систем для надежной классификации СО.

Продолжая обзор, заметим, что известны также и подходы [4,5], в которых предлагаются методы фонетического декодирования РС по его видеограммам. Эти работы интересны тем, что в них наряду с формантными траекториями используется информация о динамике интенсивности спектральных составляющих РС и предпринята попытка структурного описания звукового состава анализируемого сигнала. Однако, введение в анализ представлений об «акустических событиях», описываемых при помощи видимых объектов, по нашему мнению, недостаточно обосновано с точки зрения закономерностей восприятия слуховых образов. Эта критическая оценка опирается на результаты наших исследований процессов слухового восприятия РС, согласно которым в слуховой системе человека производится формирование и классификация СО на основе анализа более сложных (по сравнению с видимыми на видеограммах) процессов, представленных в речевом сигнале в неявном виде. Наши же результаты свидетельствуют и о неточности одного из законов восприятия РС, приведенного в [6]. В соответствии с этим законом, текущие энергетические спектры РС определяют все его психоакустические характеристики, и эти последние ни от чего

больше не зависят. Мы же установили, что в самом РС не всегда содержится информация, достаточная для формирования в СС «запланированного» при речеобразовании слухового образа, и дополнительная информация может извлекаться слуховой системой на основе анализа акустического фона среды и текущего состояния самой слуховой системы.

Одним из перспективных для разработки систем классификации СО является бионический подход. Этот подход базируется на некоторых общих принципах обработки сенсорной информации, предположительно используемых человеком. Но известные нам попытки (например, [7,8]) реализации бионического подхода ведущими специалистами по распознаванию РС [9] в целом оцениваются как малоконструктивные. Из обзора [10] также следует, что не дали позитивных результатов и работы, в которых в качестве первичного описания были использованы «слуховые спектры».

В Иркутском ВЦ СО РАН в течение нескольких лет (1992—1997 гг.)

проводились исследования, направленные на поиск «механизма»

слухового анализа речи как основы для создания дикторо—независимой

системы классификации СО речевого сигнала. В результате удалось получить ряд новых результатов. Исследования носили междисциплинарный и эволюционный характер, замыкаясь через процедуры сравнительного анализа теоретических выводов и экспериментальных результатов в итерационный цикл: «рабочая» гипотеза — функциональная модель — анализ результатов моделирования — экспериментальное выявление новых свойств системы речевой коммуникации — уточнение модели. Некоторые результаты работы представлены ниже.

О роли подсознания в восприятии слуховых образов.

Одна из основных трудностей при анализе восприятия СО состоит в неопределенности временной организации озвученных звукотипов РС. Эту задачу, как правило, решают, опираясь на известное представление о существовании в РС « квазистационарных» участков. И при решении обычно используют различные варианты методов динамического программирования или марковское моделирование. Мы для решения этой задачи применили способ анализа РС при помощи расширяющегося временного окна. Суть способа состоит в следующем.

В компьютер вводился с микрофона РС, соответствующий восприятию в

СС слова или предложения устной речи. Затем введенный сигнал обрабатывался на программном уровне и выводился на громкоговоритель. Обработка в компьютере заключалась в последовательном выделении из речи при помощи расширяющегося временного окна запланированных для анализа участков РС. Звуковой состав выделенных фрагментов РС оценивали на слух несколько аудиторов.

В качестве примера можно привести результаты оценки аудиторами слухового ощущения звукотипа /О/ из слова «кот», произнесенного мужчиной. На начальном участке РС уверенно воспринимался звукотип /Ы/. По мере увеличения длительности анализируемого РС, начинал восприниматься неопределенный звук (/Ы/ или /У/), который постепенно переходил в звукотип /У/. При дальнейшем увеличении длительности временной выборки воспринимался промежуточный звук (между /У/ и /О/). И только при достижении некоторой критической длительности прослушиваемого фрагмента начинал восприниматься звукотип /О/. Из приведенного примера следует, что слуховому ощущению (слуховому восприятию на осознаваемом человеком уровне) основного звукотипа /О/ предшествовала последовательность подсознательных процессов слухового восприятия побочных звукотипов /Ы/ и /У/, разделенных переходами от восприятия текущего звукотипа к восприятию последующего.

Исследовано более 300 вариантов озвученных звукотипов /У/, /О/, /А/, /Э/, /Ы/, /И/, выделенных из слитной и дискретной речи 19 дикторов (11 мужчин, 4 женщины и 4 детей). Звукотипы выделялись как из сильных (ударных) позиций, так и из слабых. Полученные в экспериментах оценки слухового восприятия характеризовались высокой воспроизводимостью. Основные выводы из проведенных исследо—ваний можно сформулировать следующим образом.

1). Сложный звуковой состав исследованных фрагментов РС является одной из закономерностей функционирования системы речевой коммуникации, которая находит непосредственное отражение в процессах восприятия слуховых образов речевых сигналов.

2). Система слухового анализа человека при восприятии озвученных звукотипов, представленных в анализируемой выборке, часть времени восприятия функционирует в «автоматическом» режиме, управляемом на подсознательном уровне. При этом слуховому восприятию основного звукотипа (слухового образа, воспринимаемого СС на осознаваемом уровне), как правило, предшествуют процессы восприятия побочных озвученных звукотипов (слуховых образов, воспринимаемых в СС на подсознательном уровне).

Замечание 1.

Выводы 1 и 2 коррелируют с точкой зрения Л.Заде, приведенной в докладе на конференции по нечетким системам (сентябрь,1996г.) [11]. Им была высказана идея о ведущей роли механизма «грануляции данных» в способности человека обрабатывать сенсорную информацию, поступающую из внешней среды. Главная особенность этого гипотетического механизма состоит в том, что человек не сразу трактует поступающую из среды в мозг информацию, а сначала на подсознательном уровне формирует «гранулы» — элементарные фрагменты информации, с которыми и оперирует подсознание человека.

3). Последовательности озвученных звукотипов, полученные способом расширяющегося временного окна, могут быть приняты в качестве эталонных при оценке качества имитационного моделирования механизма слухового анализа гласных устной речи. Эти последовательности содержат в явном виде информацию о временной области, соответствующей переключению СС из «режима» восприятия на подсознательном уровне в «режим» слухового ощущения, осознаваемого человеком. На этом временном интервале, вероятно, и принимается классификационное решение о воспринимаемом звукотипе.

Замечание 2.

При традиционных подходах к пофонемному распознаванию РС, выбор моментов времени , соответствующих завершению формирования эталона фонемы, обычно производится, исходя из предположения о существовании в РС упомянутых ранее квазистационарных участков. Сравнение этих значений и соответствующих значений времени, полученных экспериментальным путем, свидетельствует о том, что они могут существенно отличаться друг от друга. Это утверждение позволяет усомниться в полезности для распознавания РС как самого представления о квазистационарных участках, так и в адекватности методов, в которых используется это представление.

Об алфавите элементарных слуховых образов русской устной речи.

Выбор алфавита классифицируемых слуховой системой СО речевого сигнала является одной из ключевых задач, определяющих методологию АРСО. Отметим, что наиболее популярной у разработчиков систем распознавания РС является идея распознавания фонем речи, а по последовательности фонем и всего речевого сообщения. Однако, такой подход игнорирует существенные различия в слуховом восприятии казалось бы сходных (близких) звуков, относимых при статистических подсчетах к одному типу — фонеме. По—видимому, более адекватный алфавит мог бы быть получен по результатам оценки слухового восприятия ограниченного набора речевых звуков, описываемых знаками фонетической транскрипции. Как известно, эти знаки позволяют человеку довольно точно воспроизвести последовательность звуков. Но и такой подход к выбору алфавита в настоящее время нельзя отнести к обоснованным из—за его недостаточной изученности.

Заметим также, что в литературе нашел отражение, в виде различных теорий, широкий спектр мнений по поводу возможности представления слогов устной речи как фонетических единиц. Мы полагаем, что одним из определяющих моментов при выборе алфавита является учет тесной взаимосвязи между образованием и восприятием РС. Известно, что эта взаимосвязь состоит в способности человека адаптивно управлять процессами речевой коммуникации. Предположительно, это управление осуществляется путем сравнения некоторого эталонного слухового ощущения , «запланированного» в головном мозге говорящего, для реализации в виде речевого сигнала, с текущим слуховым ощущением в собственной СС. По результатам сравнений, вероятно, и вырабатываются управляющие воздействия на систему речеобразования, имеющие целью достижение максимально возможного сходства между эталонным и текущим слуховым ощущением. Таким образом, при речеобразовании в слуховой системе диктора производится классификация СО, которую можно рассматривать как результат принятия СС решения об удовлетворительном сходстве между эталонным и текущим слуховыми ощущениями. Учтем также и известный факт, свидетельствующий о том, что для человека «минимальной произносительной единицей устной речи является слог,представляющий собой единство слогообразующего гласного и одного или более согласных» [12]. Исходя из названных причин, естественно предположить, что слог речи является не только «минимальной произносительной», но и «минимальной классифицируемой единицей речи» , и восприятию слога можно поставить в соответствие процесс слухового восприятия некоторого элементарного слухового образа. Опираясь на это предположение, введем следующее определение. Под элементарными слуховыми образами языка устной речи понимается некоторое ограниченное множество инвариантных по отношению к ощущениям высоты и громкости эталонных слуховых ощущений речи, классифицируемых в СС в виде одного СО. Ниже приведен один из вариантов экспериментального обоснования этого представления.

Были проведены исследования субъективного восприятия набора слогов. Известно, что 93% слогов устной русской речи представлены не более, чем 3 звукотипами. На начальной стадии работы были выбраны слоги типа Г и СГ, где Г,С — соответственно, гласный и согласный устной речи. Кроме того,

исследовались и варианты сочетания гласных ГГ (например, «УО», «ОУ», «АУ», «УА» и другие). Метод психоакустических исследований выбранных слогов и набор слогов типа Г были приведены выше. Согласные звуки в слогах типа СГ были представлены сонантами /М/, /Н/, /Л/, /Р/, /Й/, а также рядом глухих и звонких шумных из РС 19 дикторов. Всего исследовано около 200 слогов типа СГ. Результаты исследований слухового восприятия выбранных РС характеризовались высокой воспроизводимостью. В итоге были установлены следующие свойства СС человека, используемые для восприятия РС.

1. Слуховое восприятие слогов типа Г характеризовалось описанными выше последовательностями побочных и основных звукотипов. Как уже отмечалось, классификация СО производилась на участке слухового восприятия основного звукотипа.

2. При прослушивании временных участков сигнала, выделенных из слогов типа СГ и соответствующих восприятию согласного, последний, как правило, воспринимался на уровне слухового ощущения, но не мог быть однозначно классифицирован в СС аудиторов. Классификационное же решение о СО принималось на этапе восприятия основного звукотипа гласного и соответствовало восприятию всего слога.
3. Слуховому восприятию сочетания гласных соответствовали два акта принятия классификационных решений в СС. Этот эффект воспроизводился независимо от порядка следования гласных в сочетании. Так, в слоге «МО» классификация производилась в момент восприятия основного звукотипа гласного, а в слове «МУО», состоящем из двух слогов и имеющем близкий к слогу «МО» состав звукотипов, однозначно фиксировались два акта классификации элементарных СО (при восприятии соответствующих гласных).

Выводы.

1. Слуховое восприятие РС, соответствующих слогам типа Г и СГ, приводило к классификации в СС только одного элементарного слухового образа.
2. Количество актов классификации элементарных СО в СС аудиторов (в исследованных фрагментах устной речи ГГ) равнялось числу слогов.

Очевидно, что различия в процессах слухового восприятия слогов типов СГ и Г от соответствующих закрытых слогов типа СГС и ГС может быть связано с присутствием в конце слогов СГС и ГС согласного. Для проверки этого предположения были проведены сравнительные экспериментальные исследования слухового восприятия слогов типа СГ и СГС, а также слогов типа Г и ГС. В результате было установлено, что процессы классификации гласных в слогах типа СГС и ГС частично идентичны соответствующим классификационным процессам при восприятии слогов типа СГ и Г. Некоторая идентичность этих процессов состояла в том, что при восприятии слогов типа Г, СГ, СГС и ГС классификация СО производилась на временных участках, соответствующих восприятию основного звукотипа гласного. Причем эксперименты показали, что в случае открытых слогов типа Г и СГ она была окончательной, а при восприятии закрытых слогов типа СГС и ГС классификация СО носила предварительный характер и в окончательной форме завершалась на участке восприятия согласного, входящего в состав слога. Таким образом, можно предположить, что такой двухэтапный механизм классификации закрытых слогов действительно существует. Этот механизм, вероятно, и позволяет различать слуховой системе элементарные СО, соответствующие восприятию слогов типа Г и ГС, содержащих одинаковые гласные. Аналогичное предположение можно сделать и о механизме классификации слогов типа СГ и СГС.

Приведенная информация носит предварительный характер. Более полное и глубокое обоснование введенных представлений мы планируем опубликовать после завершения детальных исследований процессов формирования и классификации слуховых образов в СС человека.

Заключение.

По результатам работы можно сделать следующий обобщающий вывод:

- в качестве объектов распознавания слуховых образов могут быть использованы элементарные СО , соответствующие слогам устной речи;
- обучение дикторо—независимой системы АРСО можно свести к формированию эталонных описаний элементарных СО русской устной речи;
- вся информация в речевом сигнале, потенциально используемая СС для формирования и классификации элементарных СО, располагается на временных участках РС, соответствующих слогам устной речи;
- процесс распознавания речевого сообщения можно представить в виде последовательности процедур классификации элементарных слуховых образов;
- задача получения описаний элементарных СО может быть адекватно разрешена лишь на основе выявления количественных и качественных закономерностей, определяющих сущность процессов формирования и классификации в СС элементарных слуховых образов.

Этот обобщающий вывод, по—видимому, может представлять интерес для специалистов в области речевой информатики и учтен авторами при разработке имитационной модели слухового анализа речи как основы для создания дикторо—независимой системы АРСО.

СПИСОК ЛИТЕРАТУРЫ

1. Кельманов А.В. О некоторых проблемах построения систем распознавания инвариантных к диктору. // Тез.докл.15 Всесоюз.шк.—семинара"Автоматическое распознавание слуховых образов".— Таллинн: ИК АН ЭССР, 1989.— С. 103—104.
2. Сорокин В.Н. Истинные и ложные цели в распознавании и синтезе речи. // Речевая информатика.— Киев: Ин—т кибернетики АН УССР, 1989.— С. 40—45.
3. Галунов В.И., Жаков М.Л. и др. Первичный анализ в системах авто—матического распознавания. // Тез. докл. 15 Всесоюз. шк.—семина—ра"Автоматическое распознавание слуховых образов".— Таллинн: ИК АН ЭССР, 1989.— С. 49—58.
4. Зиновьева Н.В. Особенности фонетической интерпретации "слепых" сонограмм при работе с временным окном. //Тез. докл.13 Всесоюз. шк.—семинара "Автоматическое распознавание слуховых образов". — Новосибирск: 1984. — Ч.2. — С. 30—32.
5. Зиновьева Н.В., Ампилова В.В. и др. Опыт определения опорных звуковых характеристик при анализе "слепых" сонограмм. // Тез. докл.13 Всесоюз. шк.—семинара "Автоматическое распознавание слуховых образов".— Новосибирск: 1984. — Ч.2. — С. 33—35.
6. Акинфиев Н.Н. Об одном психоакустическом законе восприятия чело—веком речевого сигнала и об объективных параметрах сигнала, содержащих речевую информацию. // Тез.докл. 15 Всесоюз. шк.—семинара "Автоматическое распознавание слуховых образов". — Таллинн: ИК АН ЭССР, 1989. — С. 179—181.
7. Галунов В.И. Бионическая модель системы распознавания речи. // Исследование моделей речеобразования и речевосприятия.— Л.: 1981, С. 36—51.
8. Галунов В.И., Родионов В.Д. Моделирование процессов передачи ин—формации в звуковом диапазоне.—Л.: 1988.—160 с.
9. Винцюк Т.К. Анализ, распознавание и интерпретация речевых сигна—лов. — Киев, Наукова думка, 1987.— 264с.
10. Люблинская В.В. Восприятие речи. Общие представления и подходы к исследованию. // Тез.докл.15 Всесоюз.шк.—семинара "Автоматичес—кое распознавание слуховых образов". — Таллинн, ИК АН ЭССР, 1989. — С.32—36.
11. Кузнецов Г. Как работает сознание. // Компьютерра, 1996, N 36.— С.17.
12. Михайлов В.Г., Златоустова Л.В. Измерение параметров речи. — М.: Радио и связь, 1987. — 168с.