О проблеме дикторонезависимости при распознавании речи на фонемном уровне.

В.В. Дубровский
Институт динамики систем и теории управления СО РАН
Россия, 664033, Иркутск, Лермонтова, 134
e-mail: arsoirkdw@mail.ru

А.И. Егоров
Институт динамики систем и теории управления СО РАН
Россия, 664033, Иркутск, Лермонтова, 134
e-mail: egorov@mail.icc.ru

Ключевые слова: дикторонезависимое распознавание речи, бионический подход, тонкая временная структура фонем, слуховая система.

Существует ли принципиальная возможность дикторонезависимого распознавания слитной, естественной речи? Ответ на этот вопрос известен: да, существует, поскольку имеется прецедент в виде человека.

Однако, есть специалисты, которые дают отрицательный ответ на поставленный вопрос. Так, в работе [1] представлены доказательства того, что человек *не* использует инвариантное к диктору описание сигнала, обеспечивающее идентификацию *всего* алфавита фонем, а, скорее всего, использует описание, которое позволяет уверенно различать лишь *группы неразличимых фонем.* Таким образом, если и существует пространство, в котором априорное распределение признаков для классов не зависит от диктора, то этими классами являются не отдельны фонемы, а группы неразличимых фонем.

Во многом соглашаясь с выводами, сформулированными в [1], авторы настоящего доклада отстаивают более оптимистичный прогноз на возможность создания неадаптивной системы распознавания речи.

Известно, что распознавание речи – многоуровневый процесс. Сначала человек воспринимает звуки, не осмысливая их, на физическом уровне. Затем идёт процесс осмысления их на основе синтаксиса, грамматики, семантики и т.д.

Зная тему разговора, часто достаточно услышать что-то весьма неопределённое, чтобы догадаться, о чём идёт речь. И наоборот: верно услышанное слово может быть принято за услышанное неверно, если оно не вяжется со смыслом или темой разговора. Всё это указывает на огромную важность уровня осмысливания во всём многоуровневом процессе распознавания речи.

С другой стороны, человек способен различать звуки, слова, фразы, которые он никогда не слышал. Здесь работает уже другой фактор

распознавания – первичный, когда слуховая система анализирует услышанное и, вероятно, получив отказ от системы осмысливания, классифицирует звуки в виде транскрипции.

Известно также, что речевой сигнал отличается от искусственных сигналов своей сложностью, неустойчивостью параметров, избыточностью. Следствием этого является то, что речевые сигналы, которые человек уверенно относит к одному и тому же образу, *никогда* не имеют полностью идентичных параметрических описаний.

Таким образом, подмена распознавания слуховых образов речи на распознавание речевого сигнала, производимое по его описаниям, при выборе которых практически полностью игнорируется участие головного мозга в порождении и восприятии речевого сигнала, скорее всего, не может привести к разработке систем для надёжной дикторонезависимой классификации слуховых образов речи.

По нашему мнению, для обеспечения дикторонезависимости более перспективным по сравнению с инженерно-кибернетическим подходом является бионический. Этот подход базируется на некоторых общих принципах обработки сенсорной информации, предположительно используемых человеком и является своеобразным вариантом признакового подхода [2].

Авторы доклада на основе известной специалистам информации из психоакустики и психофизиологии сенсорных систем [3-7], результатов собственных исследований развили названный подход и разработали бионическую систему распознавания речи.

Исследования носили междисциплинарный характер, замыкаясь через процедуры сравнительного анализа теоретических выводов и экспериментальных результатов в итерационный цикл: гипотеза – функциональная модель – анализ результатов моделирования – экспериментальное выявление новых свойств слуховой системы – уточнение модели и т. д.

В качестве элемента анализа была выбрана фонема.

В ходе исследований получены следующие результаты.

1. Показано, что определяющим фактором в формировании ощущения гласных фонем являются быстрорастущие (по сравнению с другими) гармоники основного тона. Большие по величине, но медленно растущие и, тем более, убывающие по величине гармоники в формировании фонемного ощущения играют второстепенную роль. Доказано, что область максимальной концентрации энергии (форманта) не всегда бывает определяющей в формировании ощущения.

2. Выдвинута гипотеза о существовании “оси фонем”. Гласные одной высоты расположены на оси фонем в определённом порядке, не пересекаясь. То есть, расположение гласных на оси фонем фиксировано, но зависит от высоты произнесения фонемы.

3. Доказано существование тонкой временной структуры фонем. Любая гласная фонема состоит из цепочки микрофонем многих типов.

4. Микрофонема, в свою очередь, также является сложным образованием, состоящим из нескольких фонемообразующих элементов. Вопрос о влиянии каждого фонемообразующего элемента на совокупное ощущение микрофонемы решается посредством понятия *маскировки биений биениями* и вытекающими из этого понятия следствиями. Каждому фонемообразующему элементу можно поставить в соответствие критерий слышимости, сравнивая которые можно сделать вывод о том, какие фонемообразующие элементы являются определяющими в формировании микрофонемы.

5. Определена физическая сущность микрофонемы как элемента идеально произнесённой фонемы.

6. Выявлено пространство признаков для классификации микрофонем– высота микрофонемы, и высота (или высоты) одного или нескольких фонемообразующих элементов.

7. Показано, что соответствие между фонемой и входящими в неё микрофонемами никогда не бывает взаимнооднозначным для разных реализаций фонемы, но подчиняется определённым правилам, которые также удалось установить экспериментальным путём.

8. Вопрос времени принятия решения о классе распознаваемой фонемы решён следующим образом. Одно из важнейших свойств слуха – принцип накопления ощущения. Исходя из этого принципа, при анализе фонемы невозможно выделить её главные и второстепенные временные участки. Все микрофонемы в цепочке в той или иной мере вносят вклад в формирование ощущения. В модели строится распределение текущего ощущения по времени. Находятся максимумы этого распределения, на основе чего делается вывод об итоговом (осознанном) ощущении фонемы.

Моделируемые при таком подходе процессы, по мнению авторов, согласуются с теми процессами, которые происходят в слуховой системе человека при распознавании речи.

В настоящее время разработан и реализован программно-аппаратный комплекс, позволяющий распознавать выделенные из произвольного контекста гласные русской речи без предварительной настройки на голос диктора.

То есть, на практике показана принципиальная возможность дикторонезависимого распознавания наиболее вариативных элементов русской речи, которыми являются гласные.

Литература

1. Кельманов А.В. О некоторых проблемах построения систем распознавания инвариантных к диктору. // Тезисы АРСО-15, Таллинн, ИК АН ЭССР, 1989, с. 103-104.

2. Куляс А.И. Пути создания многодикторных кооперативных систем распознавания речи // Распознавание и синтез звуковых сигналов: Сборник научных трудов, Киев, ИК АН УССР, 1987. – с. 16-24.
3. Колоколов А.С., Янко В.П. Дикторонезависимое распознавание изолированных речевых команд на основе слуховых моделей. // Автоматика и телемеханика, № 8, 1995, с. 150-157.
4. Кириллов С.Н., Стукалов Д.Н. Анализ речевых сигналов на основе акустической модели. // Техническая кибернетика, 1994, № 2, с. 147-153.
5. Гершуни Г.В. О механизме слуха (в связи с исследованием временных и временно-частотных характеристик слуховой системы). // Механизмы слуха. Л.: Наука, 1967, с. 3-32.
6. Радионова Е.А. Функциональная характеристика нейронов кохлеарных ядер и слуховая функция. // Л.: Наука, 1971.
7. Чистович Л.А., Венцов А.В., Гранстрем М.П. и др. Физиология речи. Восприятие речи человеком. // В серии «Руководство по физиологии», Л.: Наука, 1976.

On the problem of speaker-independence at recognition on the phoneme level.

W.W.Dubrowsky, A.I.Egorov

Key words: speakerindependent recognition of speech, bionic approach, complex structure of phonemes, chain of microphonemes.

In the report principles of speakerindependent recognition of speech on the base of bionic approach are stated. Known information from psychoacoustics and psychophysiology, results of our investigations are used. Major of them are the following conclusion.

1. By the determining factor in the formation of sensation of vowels is quickly growing harmonics of basic tone. The area of the maximal energy is determining factor in formation of sensation not in all cases.

2. Phonemes have complex structure. Anyone vowels consists of a chain of microphonemes of different types.

3. Conformity between vowel and microphonemes, included in it, never happens mutually unequivocal in the case of different realizations of phoneme. However, the certain rules are determined, which establish this conformity.

At present, our system of recognition allows to recognize vowels of Russian speech without any adjustments to the speaker's voice.