АВТОМАТИЧЕСКОЕ РАСПОЗНАВАНИЕ СПОНТАННОЙ УКРАИНСКОЙ РЕЧИ (НА МАТЕРИАЛЕ АКУСТИЧЕСКОГО КОРПУСА УКРАИНСКОЙ ЭФИРНОЙ РЕЧИ)

Т. В. Людовик (tetyana.lyudovyk@gmail.com)

В. В. Пилипенко (valeriy.pylypenko@gmail.com)

В. В. Робейко (valya.robeiko@gmail.com)

Международный научно-учебный центр
информационных технологий и систем
НАН Украины и МОН Украины, Киев, Украина

Работа посвящена исследованию особенностей распознавания спонтанной речи. Основное внимание уделено настройке (обучению) акустической и лингвистической моделей, а также словарю словоформ с транскрипциями, отражающими спонтанное произнесение.

Ключевые слова: спонтанная речь, распознавание речи, акустический корпус, спонтанное произнесение.

AUTOMATIC RECOGNITION OF SPONTANEOUS UKRAINIAN SPEECH BASED ON THE UKRAINIAN BROADCAST SPEECH CORPUS

T. V. Liudovyk (tetyana.lyudovyk@gmail.com)

V. V. Pylypenko (valeriy.pylypenko@gmail.com)

V. V. Robeiko (valya.robeiko@gmail.com)

RAS of Ucraine, Kiev, Ucraine

The paper focuses on automatic recognition of spontaneous Ukrainian speech, introducing the Acoustic Corpus of Ukrainian Media Speech (ACUMS) Three confi gurations of a speech recognition system are considered. Special attention is paid to training basic and thematic acoustic and linguistic models as well as to the lexicon that contains word transcriptions

refl ecting spontaneous pronunciation. The basic acoustic model was trained on recordings from approximately 2,000 speakers (52 hours). The basic language model was trained on ACUMS texts and on texts taken from Internet (400 Mb). Spontaneous variants of word transcriptions were obtained automatically based on standard Ukrainian pronunciation. Experimental results show that clear normative speech is recognized 50 % better than less intelligible speech with hesitations and reductions. Errors are due mainly to erroneous speech corpus annotation, non-vocabulary words (proper names in particular), spontaneous manner of pronunciation, short reduced words (conjunctions and prepositions), and a strong impact of language model on the algorithm searching for the best word sequence.

Key words: spontaneous speech, speech recognition, speech corpus, spontaneous pronunciation.

1. Введение

Многие прикладные задачи в области речевых технологий связаны с распознаванием спонтанной речи. Однако, точность ее распознавания, достигаемая современными системами распознавания речи (automatic speech recognition, ASR-системами), далека от точности распознавания подготовленной (прочитанной) речи, а тем более от точности распознавания изолированно произносимых слов.

Еще в 90-х годах в [1] было проведено сравнение распознавания спонтанной и подготовленной речи на материале одних и тех же 20 дикторов и одних и тех же текстов (сначала дикторы спонтанно вели диалоги, а затем читали тексты-записи своей спонтанной речи). Этот эксперимент показал, что стиль речи является главным фактором, влияющим на точность распознавания: пословная точность распознавания прочитанных текстов составила 62,4%, тогда как точность распознавания разговорной речи — всего 47,4%.

Спонтанная речь труднее поддается автоматическому распознаванию в первую очередь из-за ее вариативности, моделированию которой уделяется большое внимание [2]. Вариативность проявляется как на аллофонном, так и на фонемном уровнях.

Особенности спонтанной речи проявляются также в нарушении плавности речи, выраженном в виде пауз хезитации (например, «а-а», «э-э»), повторов слов или их начальных частей, оговорок, а также нарушение синтаксического оформления высказывания. Не менее важен характер лексики спонтанной речи (использование социальных диалектов, «суржика»).

Следует подчеркнуть, что у разных дикторов спонтанная речь характеризуется разными особенностями: одним мало свойственны паузы хезитации (дипломаты, актеры), другим свойственно хезитативное удлинение, третьим — редуцированное произношение. Широко применяемые в распознавании речи статистические методы «усредняют» дикторов.

Как правило, в ASR-системах используются методы, основанные на скрытых Марковских моделях (hidden Markov models (HMMs) [1, 3]).

Акустические модели фонем, составляющие общую акустическую модель (АМ), получаются путем предварительного обучения системы на большом массиве данных, включающем несколько десятков или сотен часов звучащей речи вместе с ее транскрипцией [4–7]. Акустические модели учитывают аллофонную вариативность произношения (в пределах фонемы). Не зависящее от диктора (дикторонезависимое, многодикторное) распознавание речи требует для обучения речь многих сотен дикторов [8].

Лингвистическая модель (ЛМ) задает возможные последовательности слов либо в явном виде, либо в виде вероятностей следования одних слов за другими. В последнем случае ЛМ получается (обучается) путем предварительного анализа большого массива текстов.

Третьим компонентом ASR-системы является словарь словоформ с транскрипциями (словарь распознавания), используемый непосредственно в процессе распознавания [7, 9]. Именно в этом словаре должна быть отражена вариативность произношения на фонемном уровне, свойственная спонтанной речи. Однако, простое расширение словаря транскрипций за счет добавления вариативных произнесений иногда приводит не к повышению, а к понижению точности распознавания из-за того, что разные слова представлены одинаковыми или похожими транскрипциями [2, 9–11]. Тем не менее, удачный выбор количества вариантов транскрипций одного слова позволил повысить точность распознавания с 78 % до 85,7 % [4].

Обученные АМ и ЛМ, а также словарь распознавания используются в процессе распознавания речи для поиска наиболее вероятной последовательности слов, соответствующей входному речевому сигналу.

В данной работе основное внимание уделено:

А) акустической модели;

Б) лингвистической модели;

В) словарю, используемому при распознавании.

Исследуется распознавание спонтанной украинской речи, в частности, применительно к конкретной предметной области (судебные заседания).

2. Цель исследования

Основной задачей исследования было повышение точности распознавания украинской спонтанной речи.

Были поставлены следующие цели:

- провести эксперимент с использованием базовой системы распознавания украинской речи;
- проанализировать ошибки, предложить и реализовать меры по повышению точности распознавания с учетом спонтанного характера речи и сужения ее тематики; провести соответствующие эксперименты;

- сравнить результаты распознавания спонтанной речи актеров в роли судьи на материале телепередач и речи реального судьи на материале выступлений в ходе судебных заседаний.

3. Материал для исследований

3.1. Речевой материал

Речевой материал для исследований был взят из Акустического корпуса украинской эфирной речи (АКУЭМ) [12]. В АКУЭМ представлена как украинская, так и русская речь, как подготовленная, так и спонтанная. Русская речь в данной работе не анализировалась.

Для экспериментов по распознаванию речи, относящейся к судебной тематике, использовалась только часть аудиофайлов. Это в основном записи телепередач «Судові справи» («Судебные дела»). Речь, звучащую в этих телепередачах, можно назвать спонтанной по форме, но не по содержанию, поскольку дикторы говорили в рамках соответствующих ролей. Кроме этого, часть аудиофайлов содержит записи реальных судебных заседаний, в которых присутствует как спонтанная речь судьи, так и неподготовленное (и, таким образом, приближенное к спонтанному) чтение протоколов.

Речевой материал, использованный для построения АМ, состоял из аудиозаписей (длительностью около 52 часов), в которых содержится речь около 2000 дикторов. Распределение неравномерное: большинство дикторов представлено короткими записями, однако, у 150 дикторов длительность записей составляет более 10 минут.

3.2. Текстовый материал

Текстовый материал, использованный для построения лингвистических моделей, состоит из текстов корпуса АКУЭМ, и текстов, загруженных из Интернета (400 Мбайт).

3.3. Контрольная выборка

Контрольная выборка для всех экспериментов использовалась одна и та же. Для распознавания использовались записи длительностью 3,74 часа, в которых встретилось 29 500 реализаций слов. Всего в контрольной выборке присутствовала речь 34 дикторов. Темп произнесения — средний и быстрый.

В таблице 1 представлены характеристики речи некоторых дикторов, чья речь вошла в контрольную выборку. Внимание было обращено на речь «главных действующих лиц» судебных заседаний: судей, прокуроров, адвокатов, судебного секретаря, судебного пристава.

Таблица 1. Характеристики речи некоторых дикторов контрольной выборки

Дикторы	Пол	Профессия	Степень нормативности	Степень разборчивости	Склонность к хезитации	Склонность к редукции
Окис	м.	актер в роли судьи	средняя	средняя	слабая	сильная
Калинская	ж.	актриса в роли судьи	средняя	средняя	слабая	очень сильная
Ш.	ж.	судья	средняя	средняя	слабая	слабая
Антонюк	ж.	актриса в роли прокурора	средняя	средняя	сильная	слабая
Наум	м.	актер в роли прокурора	низкая	низкая	сильная	сильная
Бойко	м.	актер в роли прокурора	низкая	средняя	сильная	сильная
Бевз	м.	актер в роли адвоката	средняя	средняя	сильная	сильная
Жуковская	ж.	актриса в роли адвоката	средняя	средняя	сильная	слабая
Бабич	ж.	актриса в роли адвоката	средняя	средняя	сильная	сильная
Бузаджи	м.	актер в роли адвоката	средняя	средняя	сильная	сильная
Солодко	м.	актер в роли адвоката	средняя	средняя	сильная	сильная
Сологуб	ж.	актриса в роли судебного секретаря	высокая	высокая	слабая	слабая

4. Система распознавания речи

Для исследований использовался инструментарий HTK [13]. На его основе была создана многодикторная система распознавания речи [5]. На рисунке 1 представлены базовая конфигурация системы распознавания речи и две ее модификации, отличающиеся комбинациями АМ, ЛМ и словарей.

4.1. Описание базового варианта системы

Предварительная обработка речевого сигнала описана в [5].

В качестве АМ используются скрытые Марковские модели, обученные на всей украинской речи корпуса АКУЭМ всех дикторов. 56 украинских контекстно-независимых фонем (включая фонему-паузу) моделируются тремя состояниями Марковской цепи без пропусков. Используется диагональный вид Гауссовских функций плотности вероятности. Редко встречающиеся фонемы моделируются 64 смесями Гауссовских функций плотности вероятности, более часто встречающиеся фонемы моделируются большим числом смесей, наиболее часто встречающиеся фонемы используют 1024 смесей.

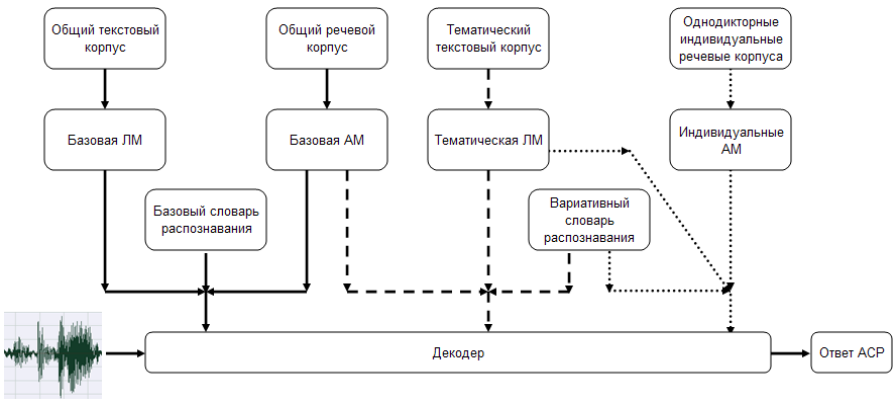

Рис. 1. Различные конфигурации системы распознавания речи: базовая (сплошные стрелки) и ее возможные комбинации: тематическая (пунктирные стрелки) и индивидуальная (точечные стрелки)

В корпусе АКУЭМ отмечены такие паралингвистические явления как вдох-выдох, кашель, смех, плач, причмокивание, а также паузы хезитации («а-а-а», «е-е-е», «м-м», …). Различаются фоновые паралингвистические явления (наложенные на речь) и изолированные. Последние при построении АМ рассматриваются как отдельные слова, состоящие из одной фонемы. Распознанные слова-»паралингвизмы» впоследствии удаляются из окончательного ответа.

Для построения ЛМ тексты, загруженные из Интернета, и тексты корпуса АКУЭМ были модифицированы с целью удаления служебной информации, записи чисел в текстовом виде, а также отделения украиноязычных фрагментов от русскоязычных. На основе полученного материала была построена биграммная модель языка, заданная вероятностями появления пар слов. Поскольку в текстах, на которых вычислялись статистики, встретились далеко не все пары слов, входящие в словарь распознавания системы, для аппроксимации ненайденных пар слов использовались обратные (back off) коэффициенты.

Словарь распознавания базовой системы насчитывает 42598 словоформ. Произнесение каждой словоформы представлено транскрипцией, несколько

отличающейся от канонической (литературной). А именно, односложные словоформы представлены двумя транскрипциями (ударный и безударный варианты), а также упрощены некоторые сочетания согласных в соответствии со спонтанным произнесением (например, «дч» → «чч» вместо канонического «джч»).

4.2. Результаты распознавания базовым вариантом системы

Одним из главных показателей работы систем автоматического распознавания речи является точность (надежность) распознавания. В проведенных экспериментах точность распознавания измеряется путем сравнения орфографических транскрипций, имеющихся в корпусе, с текстами, получаемыми на выходе системы распознавания речи.

Обычно измеряется пословная точность распознавания. Ошибками считаются вставки, пропуски, замены слов. Вставки, пропуски и замены на уровне лексем являются серьезными, поскольку не позволяют восстановить смысл произнесенной фразы. Замены, вставки и пропуски на уровне словоформ являются незначительными, поскольку, как правило, не искажают смысл произнесенной фразы, а лишь нарушают ее грамматическое оформление.

Таблица 2. Результаты распознавания речи контрольной выборки базовым вариантом системы

Дикторы	Профессия	Точность распознавания (%)
Окис	актер в роли судьи	73,47
Калинская	актриса в роли судьи	58,65
Ш.	судья	59,47
Антонюк	актриса в роли прокурора	63,90
Наум	актер в роли прокурора	59,10
Бойко	актер в роли прокурора	57,76
Бевз	актер в роли адвоката	55,93
Жуковская	актриса в роли адвоката	66,38
Бабич	актриса в роли адвоката	51,64
Бузаджи	актер в роли адвоката	60,28
Солодко	актер в роли адвоката	46,95
Сологуб	актриса в роли судебного секретаря	81,26

В таблице 2 представлены результаты распознавания речи контрольной выборки базовым вариантом системы.

Ниже приведены примеры в виде пар «произнесено диктором — распознано системой».

Диктор-мужчина — судья Окис:

А) произнесено: «***введення наркотичних засобів*** *в організм* ***людини***»

Б) распознано: «***введення наркотичних засобів*** *організму* ***людини***»

Диктор-мужчина — адвокат Бевз:
А) произнесено: «*то* ***що ми*** *передивилися* ***файл***»
Б) распознано: «*тому* ***що ми*** *подивилися* ***файл***»

4.3. Анализ ошибок распознавания базовой системой

Обнаруженные ошибки можно разделить на несколько групп:

1) Ошибки, допущенные стенографистами и экспертами на этапе аннотирования корпуса АКУЭМ и словаря распознавания. Обычно это орфографические ошибки, а также необозначение таких явлений как звучащие паузы, вдох, смех, оговорки, фоновые звуки (например, музыка) и речь посторонних дикторов. Ошибки, допущенные на этапе аннотирования корпуса, приводят в дальнейшем к ошибкам распознавания. Примеры: «*слухається*» (правильно «*слухається*»), «*пятнадцять*» (правильно «*п'ятнадцять*»).
2) Ошибки, связанные со словарем распознавания. Это или отсутствие в словаре слов, содержащихся в речи (OOV — out of vocabulary ошибки), или ошибки в орфографии и/или транскрипции. Наиболее часто в разряд OOV попадают фамилии и географические названия. Примеры:фамилия «*гуріна*» распозналась как «*горі на*», фамилия «*фещук*» распозналась как «*те що*».
3) Ошибки, связанные с вариативностью спонтанного произношения. В спонтанной речи некоторые слова теряют ударение, наблюдается редукция (сокращение, упрощение произнесения и выпадение отдельных звуков и звукосочетаний). Как правило, сильно редуцируются при произнесении часто употребляемые слова и выражения. Например, «*слово*» было произнесено как [слО] и распозналось как «*село*» (по-украински произносится «*сэло*»). Аналогично, «*як ви бачите*» [йакубАчити] распознналось как «*я побачити*», «*знаємо*» [знАйми] распозналось как «*з нами*».
4) Ошибки, связанные с распознаванием коротких слов (предлогов, союзов), присутствие которых в речи скорее «угадывается», чем слышится на самом деле.
5) Ошибки, связанные с алгоритмом поиска правильной последовательности слов. Правильная гипотеза может быть отброшена из-за ее малой вероятности, предсказываемой ЛМ.

Высокая степень нормативности и разборчивости речи, а также отсутствие склонности к хезитации и редукции положительно сказываются на точности распознавания (в среднем 81,26% точности). Речь с низкой степенью нормативности и разборчивости распознается плохо (58–59%). Склонность диктора одновременно к хезитации и редукции отрицательно сказывается на точности распознавания (47–60%).

Результаты распознавания речи базовой системой показали, что речь актеров, исполняющих в телепередачах роли судей, прокуроров и адвокатов, распознается лучше, чем речь реального судьи, записанная во время судебного заседания. Это может быть объяснено тем, что в реальных обстоятельствах средний темп речи более быстрый.

5. Влияние тематической (судебной) ЛМ и индивидуальных АМ на точность распознавания речи

Целью экспериментов была проверка того, как влияют на точность распознавания речи а) ограничение тематики; б) максимальное ограничение количества дикторов. В последнем случае речь идет фактически об однодикторной системе распознавания речи.

5.1. Тематическая (судебная) ЛМ

ЛМ, ориентированная на судебную тематику, строилась на текстах из трех источников:

1) Тексты из Интернета (400М);
2) Тексты на судебную тематику, выделенные из АКУЭМ путем автоматической кластеризации;
3) Искусственно созданный текст, в который вошли, в частности, последовательности числительных и обозначения дат (например, «тридцяте січня дві тисячі одинадцятого року»).

5.2. Индивидуальные АМ

Были созданы две индивидуальные АМ на основе речи отдельных дикторов. Для одной из них обучающий материал составил 1,91 часа речи актера-мужчины, исполняющего роль судьи. Вторая индивидуальная модель была обучена на речи диктора-женщины, играющей роль прокурора (1,4 часа).

5.3. Словарь распознаваниия, учитывающий спонтанное произнесение

Все слова можно условно разделить на классы, встречающиеся в речи с разной частотой и подвергающиеся разной степени редукции [9]. В связи с этим было произведено разбиение общего словаря, используемого на этапе распознавания, на подсловари, отличающиеся количеством вариантов произнесения, приходящихся на одну словоформу.

Наибольшее число вариантов транскрипций имеют подсловари «наиболее частотные общеупотребительные слова» (1000 словоформ), «наиболее частотные слова судебной тематики» (1000 словоформ, без пересечения с общеупотребительными), «имена, отчества и фамилии», а также «числительные».

Меньшим количеством вариантов транскрипций представлены «имена собственные географические», «аббревиатуры», «социальные и территориальные диалекты, «суржик» и «устойчивые словосочетания».

Таким образом, более часто встречающиеся словоформы представлены большим числом транскрипций. Большая часть дополнительных транскрипций была получена автоматически с использованием правил, выведенных на основе анализа речевого материала [5]. Меньшая часть дополнительных транскрипций была написана вручную.

Приравнивание устойчивых словосочетаний к отдельным словоформам («ваша честь», «будь ласка») позволяет частично учитывать фонемную и аллофонную вариативность на стыках словоформ.

Объем тематического словаря, учитывающего вариативность спонтанного произнесения, составляет 22 947 словоформ, в среднем 1,35 транскрипций на одну словоформу. В таблице 3 приведены примеры словоформ и их транскрипций.

Таблица 3. Примеры вариантов транскрипций словоформ

Словоформа	**Литературная фонемная транскрипция**	**Спонтанные фонемные транскрипции**
виявлено	в И й а в л е н о	в И й а л е н о в И й л е н о в И й л и н и
п'ятнадцять	п й а т н А дз' ц' а т'	п й а т н А ц' а т' п' а т н А ц' а т' п' а т н А ц'
в'ячеславовича	в й а ч е с л А в о в и ч а	в' а ч е с л А в о в и ч а в' а ч е с л А в и ч а

Результаты распознавания контрольной выборки конфигурацией системы, включающей базовую АМ, судебную ЛМ и вариативный словарь, ориентированный на спонтанное произнесение (рис. 1) в целом не отличаются от результатов, достигнутых базовым вариантом системы.

Конфигурация из индивидуальной АМ, судебной ЛМ и вариативного словаря, как и следовало ожидать, показала лучшие результаты, чем базовая модель. Обучение АМ только на речи актера, исполняющего роль судьи, повысило точность распознавания его речи на 3 % (с 73,47 % до 76,84 %). Обучение АМ только на речи актрисы, играющей роль прокурора, привело к повышению точности распознавания ее речи на 5 % (с 63,90 % до 69,24 %).

6. Направления будущих исследований

Для повышения точности распознавания спонтанной речи многодикторной системой представляются целесообразными исследования в следующих направлениях:

- Совершенствовать АМ путем адаптции к отдельным дикторам или группам дикторов, а также к темпу речи.

- В случае отсутствия распознаваемого слова в словаре выдавать его фонетическую транскрипцию.
- При различении омофонов учитывать их частотность в конкретной предметной области.
- Сбалансировать набор правил, порождающих варианты транскрипций спонтанного произнесения, и адаптировать эти правила к произношению отдельных дикторов.
- Автоматизировать выявление часто встречающихся устойчивых словосочетаний (multi-words) в рамках предметных областей.
- Предложить более гибкий критерий оценки правильности распознавания, в частности, замены одной словоформы лексемы другой ее словоформой считать менее грубой ошибкой, чем замену одной лексемы другой лексемой.

7. Выводы

Базовая система распознавания речи обеспечивает 59,61 % точности распознавания спонтанной украинской речи. Наилучшие результаты достигнуты при распознавании речи диктора, отличающегося высокой степенью нормативности произношения, отсутствием склонности к хезитации и редукции.

Ограничение тематики распознаваемой спонтанной речи не привело к повышению точности.

Индивидуальные АМ позволили значительно (на 3–5 %) повысить точность распознавания.

Система распознавания спонтанной речи может быть использована для автоматизации документооборота в судах.

References

1. *Amdal I., Fosler-Lussier E.* 2003. Pronunciation Variation Modeling in Automatic Speech Recognition. Telektronikk : 70–82.
2. *Burdic J.* 2004. Building a Regionally Inclusive Dictionary for Speech Recognition. http://surj.stanford.edu/2004/pdfs/burdick.pdf
3. *Byrne B., Finke M., Khudanpur S., McDonough J., Nock H., Riley, M., Saraclar M., Wooters C., Zavaliagkos G.* 1997. Pronunciation Modeling for Conversational Speech Recognition: a Status Report from WS97. Automatic Speech Recognition and Understanding :26–33.
4. *Hillard D., Hwang M., Harper M., Ostendorf M.* 2008. Parsing-Based Objective Functions For Speech Recognition In Translation Applications. ICASSP : 5109–5112.
5. *Niklolenko S. I., Korenevskii M. L., Ponomareva I. A., Levin K. E.* 2010. Double Recognition Based on Speech Thematic Classification. Chetvertyi Mezhdistsiplinarnyi Seminar “Analiz Razgovornoi Russkoi Rechi” : 28–32.

6. *Ostendorf M., Byrne B., Bacchiani M., Finke M., Gunawardana A., Ross K., Roweis S., Shriberg E., Talkin D., Waibel A., Wheatley B., Zeppenfeld T.* 1996. Modeling Systematic Variations in Pronunciation Via a Language-Dependent Hidden Speaking Mode. Proc. Intl. Conf. on Spoken Language Processing.
7. *Pilipenko V. V., Robeiko V. V.* 2008. Automatic Stenographer of Ucrainian Speech [Avtomatizirovannyi Stenograf Ukrainskoi Rechi]. Iskustvennyi Intellekt, 4 : 768–775.
8. *Rabiner L. R., Juang B. H.* 1986. An Introduction to Hidden Markov Models. IEEE ASSP Mag : 4–16.
9. *Strik H., Cucchiarini C.* 1998. Modeling Pronunciation Variation for ASR: Overview and Comparison of Methods. Proc. ESCA Workshop on Modeling Pronunciation Variation for Automatic Speech Recognition:137–144.
10. *Vasil'eva N. B., Pylypenko V. V., Raduts'kii O. M., Robeiko V. V., Sazhok M. M.* 2010. Creation of the Acustic Spoken Ucrainian Speech Corpus [Stvorennia Akustichnogo Korpusu Ukrains'kogo Efirnogo Movlennia]. Obrabotka Signaliv Izobrazhen' ta Rospiznavannia Obraziv: 10 Vseukrains'ka Mizhnarodna Konferentsiia : 55–58.
11. *Weintraub M., Taussig K., Hunicke-Smith K., Snodgrass A.* 1996. Effect of Speaking Style on LVCSR Performance. Proc. Intl. Conf. on Spoken Language Processing. Philadelphia :16–19.
12. *Young S.* et al. 2009. The HTK Book (for HTK Version 3.4), available at: http://htk.eng.cam.ac.uk/
13. *Zulkarneev M. Iu., Satunovskii P. S.* 2009. Variability of Pronunciation Modeling using Hidden Markov Models [Modelirovanie Variativnosti Proiznosheniia s Ispol'zovaniem Skrytykh Markovskikh Modelei]. Tretii Mezhdistsiplinarnyi Seminar "Analiz Razgovornoi Russkoi Rechi" : 74–78.