УДК 004.934

Н.Б. Васильева, В.В. Робейко, Д.Я. Федорин

Эффективность реализации кросс-платформенных систем распознавания речи

Статья посвящена поиску наиболее эффективной реализации систем распознавания речи на различных вычислительных платформах и формированию базы данных и знаний акустического, фонетического и лексического уровней. Моделируется связь акустической и лингвистической компонент системы распознавания речевого сигнала, исследуется эффективность выбора речевых элементов. Описаны особенности реализации системы распознавания на архитектуре микропроцессоров цифровой обработки сигналов и возможность удаленной обработки речевого сигнала.

The paper is devoted to finding the most effective speech recognition system implementation for a variety of computing platforms. Particular attention is given to the data and knowledge base forming for acoustic, phonetic and lexical levels. Relation between speech recognition acoustic and linguistic components is being modeled as well as spoken element selection has been investigated. Aspects of decoder implementation on the DSP microprocessor architecture including the possibility of speech signal remote processing are described.

Статтю присвячено пошуку найбільш ефективної реалізації систем розпізнавання мовлення на різних обчислювальних платформах та формуванню бази даних і знань акустичного, фонетичного та лексичного рівнів. Моделюється зв'язок акустичної та лінгвістичної компонент системи розпізнавання мовленнєвого сигналу, досліджується ефективність вибору мовленнєвих елементів. Описано особливості реалізації системи розпізнавання на архітектурі мікропроцесорів цифрового оброблення сигналів і можливість віддаленої обробки мовленнєвого сигналу.

Введение. Распознавание слитной речи в реальном времени позволяет решать широкий спектр прикладных задач в различных областях человеческой жизни. Анализ патентов коммерческих фирм и публикаций известных научных центров мира показывает, что в последнее десятилетие появилось много программных средств диктования на ПК, а также сетевые сервисы, позволяющие устно формировать поисковые запросы или диктовать письма электронной почты. Все наиболее продуктивные системы реализуют генеративную модель анализа, распознавания и понимания речевого сигнала в той или иной модификации [1–3]. Эффективность приложений систем распознавания речи зависит от настраивания параметров множества компонент, что до сих пор не упорядочено. Остаются открытыми вопросы выбора элементов распознавания на разных уровнях и взаимосвязи уровней системы распознавания.

Реализация алгоритмов распознавания и синтеза речи в портативных устройствах – чрезвычайно актуальная проблема. Прежде всего это касается алгоритма распознавания больших словарей, т.е. пофонемного распознавания отдельно произносимых слов, причем количество слов, которые система может распознать (словарь), составляет 1000 элементов и более. Еще более важны проблемы распознавания слитной спонтанной речи (распознавание последовательности слов, произнесенных диктором без предварительной подготовки) и синтеза на естественном языке произвольного текста. Эти алгоритмы можно использовать для управления портативными устройствами, перевода сказанного на другие языки, голосового поиска, построения диалоговых систем и при решении множества других задач.

В зависимости от места, где происходит превращение «произнесенная фраза – текст» и «текст – произнесенная фраза» программы распознавания и синтеза речи делятся на *изолированные* (*client-side*), *клиент-серверные* (*server-side*) и *гибридные* (*hybrid*). В изолированных системах все преобразования происходят непосредственно на мобильном устройстве. В клиент-серверных – мобильное устройство используется только для ввода информации и передачи ее по сети на сервер для дальнейшей обработки и получения от сервера ответа распознавания или синтезированной фразы. *Гибридные* системы совмещают функциональность изолированных и клиент-серверных, при наличии доступа к сети они используют для преобразования сервер, при недоступности – работают как изолированная система. Примером реализации изолированной системы может быть си-

стема *CeedVocal* [4], примером клиент-серверной – общеизвестная *Siri* [5], примером гибридной – *VoCon Hybrid* [6]. Каждый из подходов имеет свои преимущества и недостатки. Изолированная система ограничена быстродействием и размером доступной оперативной памяти современных мобильных систем, что в свою очередь накладывает ограничения на размер словаря и продлевает время ответа приложения. Клиент-серверная технология не имеет этих ограничений, но нуждается в постоянном подключении к глобальной сети. Гибридная технология, будучи, по сути, реализацией двух предыдущих технологий в одной системе, – наиболее гибкая.

Далее рассмотрим общую структуру распознавания речи, проанализируем эффективность ее компонент в отдельности и во взаимосвязи, а также рассмотрим особенности адаптации речевых систем к различным, в особенности, к портативным платформам.

Общая структура системы распознавания речи

Входящий речевой сигнал преобразуется в последовательность акустических векторов–признаков $\boldsymbol{Y}_{1:T} = (\boldsymbol{y}_1, \boldsymbol{y}_2, ..., \boldsymbol{y}_T)$ в результате препроцессинга. Затем декодер пытается отыскать последовательность речевых сегментов, заданных языковыми символами, $\mathbf{w}_{1:L} = (w_1, w_2, ..., w_L)$, которая наиболее вероятно соответствует наблюдаемому $\boldsymbol{Y}$:

$$\hat{\mathbf{w}} = \underset{\mathbf{w}}{\operatorname{argmax}} P(\mathbf{w} \mid \mathbf{Y}) \cong \underset{\mathbf{w}}{\operatorname{argmax}}\, p(\mathbf{Y} \mid \mathbf{w}) P(\mathbf{w}). \quad (1)$$

Эквивалентность правой части выражения, вытекающая из применения правила Байеса, представляет собой базовую формулировку генеративной модели распознавания речи. Акустическая – $p(\boldsymbol{Y} \mid \boldsymbol{w})$ и лингвистическая – $P(\boldsymbol{w})$ составляющие генеративной модели описываются каждая своими стохастическими порождающими грамматиками.

Акустическая модель каждого из слов w формируется в результате композиции моделей базовых речевых элементов, т.е. фонем, составляющих фонемную транскрипцию слова $\mathbf{q}^{(w)}_{1:K_w} = (q_1, q_2, ..., q_{K_w})$. Для моделирования экстралингвистических явлений, свойственных спонтанной речи, в алфавит базовых элементов, дополнительно к фонемам и фонемам–паузам, вводятся символы, отображающие неинформативные звуки.

Общепринятые системы пофонемного распознавания оперируют алфавитом фонем контекстно-зависимых или контекстно-независимых, из которых строятся речевые образы слов. Уже на последовательности слов накладываются ограничения путем введения лингвистической модели на основе порождающих грамматик или статической модели, учитывающей контексты слов.

На рис. 1 изображена общая структура системы автоматического распознавания речи в реальном времени.

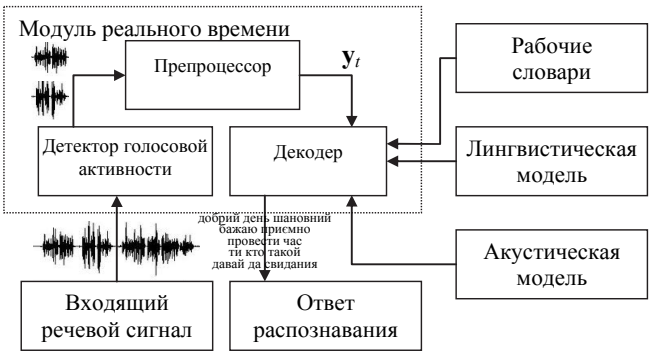

Рис. 1. Структура автоматического распознавателя спонтанной речи

В *модуль реального времени* поступает *речевой сигнал* через один из доступных источников (микрофон или файл). При прохождении через *детектор голосовой активности* сигнал разбивается на сегменты по признаку наличия голосового ввода. Используются простые акустические признаки в амплитудно-временном пространстве на основе текущей амплитуды и количества переходов через ноль. *Блок препроцессора* переводит сигнал в пространство первичных векторов–признаков. При этом используется мел-кепстральное преобразование с вычитанием среднего значения. *Декодер* сравнивает входящий речевой сегмент с гипотезами эталонного сигнала допустимых последовательностей словарных сегментов из *рабочих слова-*

рей, применяя некую осторожную стратегию отбрасывания малоперспективных гипотез. Для этого используются данные из *акустической* и *лингвистической* моделей. Последовательность слов, по которой генерируется эталонный сигнал, наиболее похожий на входящий сигнал, объявляется *ответом распознавания*.

Формирование баз данных и знаний для систем распознавания речи

База данных и знаний для системы распознавания речи включают в себя рабочие словари, акустическую и лингвистическую модели.

В *рабочем словаре* содержатся варианты произношений для каждого из слов предполагаемого лексикона. Графемно-фонемные преобразования нужны для формирования словарей произношения при оценке параметров акустической модели. Именно в этих словарях должна быть отражена вариативность произношения на фонемном уровне, свойственная спонтанной речи. В ходе работы по распознаванию речевого сигнала на основе анализа спонтанной речи нескольких сотен дикторов была разработана система, на вход которой подается орфографический текст, содержащий только символы из алфавита букв, включая символы границы между словами и морфемами и обозначением ударения. На выходе системы получаем последовательности фонем, соответствующие различным вариантам произношения входящего текста [7]. Разработанная система многозначного транскрибирования орфографических текстов использует конечный автомат, предусматривающий возможность таблично задавать контекстно-зависимые правила преобразований одних обобщенных последовательностей символов в другие. При этом в каждом правиле задается ширина шага, по которому происходит переход к следующей последовательности символов. Для построения базовой транскрипции украиноязычных текстов достаточно не более 30–35 правил. Применение многих правил позволяет генерировать сразу несколько вариантов транскрипции одного и того же слова или нужный вариант из нескольких возможных, например, описывая спонтанную речь одного диктора или группы.

Возможность генерировать сразу несколько вариантов транскрипции одного и того же слова позволяет продемонстрировать в словаре вариативность произношения наиболее частотных украинских слов, редуцирование и растяжение слов при быстром темпе речи, нечеткое произношение и подобные явления наряду с литературным вариантом произношения. Также система транскрибирования позволяет генерировать транскрипции для таких специфических подсловарей, как словарь суржика, социальных и территориальных диалектов, аббревиатур и пр.

Спонтанная речь характеризируется динамичностью на лексическом уровне. Постоянно появляются новые слова и выражения, интенсивно используется диалектная и «суржиковая» лексика, обсценная лексика. В существующих используемых словарях ударений [8] невозможно зафиксировать и передать многообразие лексических форм, а привлечение экспертов для составления дополнительных словарей ударений предусматривает значительные трудовые затраты. Предлагается использовать алгоритм, в котором решение о месте ударения в слове принимается на основе знаний об ударениях в оговоренном словаре ударений и с использованием массива текстов [9].

Связь словаря с акустической и лингвистической моделями осуществляется по идентификатору (имени), дополненному вероятностью принадлежности к кластеру слов для моделей, основанных на классах слов [10].

Параметры акустической модели оцениваются на основании *речевого корпуса*, состоящего из структурированного множества речевых фрагментов, описания этих фрагментов, а также инструментария для оперирования со всем множеством данных корпуса.

При создания речевого корпуса с последующим формированием обучающих и контрольных выборок исходим из первичности описания речевых фрагментов. Это позволяет избежать этапа ручного транскрибирования и сегментирования, а также одновременно формировать *текстовый корпус* (ТК), соответствующий исследуемой предметной области. Пред-

полагается, что диктор зачитывает текст, в котором содержится все фонетическое разнообразие украинской речи. Формирование такого текста происходит на базе текстов, наличных в свободном доступе в Интернете.

В процессе формирования текста *обучающей выборки* (ОВ) *слитной речи* проводилось преобразование цифр, символов и сокращений в последовательность графем, которые преобразовались в фонемы [7] и, наконец, применялся «жадный» алгоритм (ЖА) [11], позволивший достичь существенного сокращения текста ОВ без потери фонемного разнообразия. На этапе обработки текстового корпуса и формирования ОВ рассматривались фонемы–трифоны как базовые речевые образы, поскольку они имеют регулярную структуру и дают возможность моделировать фонемное разнообразие, учитывая правый и левый звуковые контексты.

Графики частотности фонем–трифонов в разных источниках (ТК, словарь УМИФ [8] и частотный словарь) и полученных соответствующих ОВ приведены на рис. 2. Здесь можно увидеть, что при работе ЖА количество элементов, встречающихся лишь однажды, увеличивается в несколько раз для каждой ОВ. Также из рисунка следует, что частоты фонем–трифонов примерно соответствуют распределению Ципфа–Мандельброта как для исходных корпусов, так и после работы ЖА.

В табл. 1 приведены статистические данные по фонемам–трифонам в ОВ до и после применения ЖА.

Для формирования ОВ из *изолированных слов* использованы частотный словарь украинского языка и словарь УМИФ [8]. Количество фонем–трифонов, принадлежащих обеим словарным выборкам, составляет приблизительно 15 тыс. элементов. При этом 12 тыс. фонем–трифонов принадлежат только ОВ на основе словаря УМИФ, а 3 тыс. – только ОВ частотного словаря. Объем словаря ОВ изолированных слов составил 13 тыс. слов (около 12 часов записи).

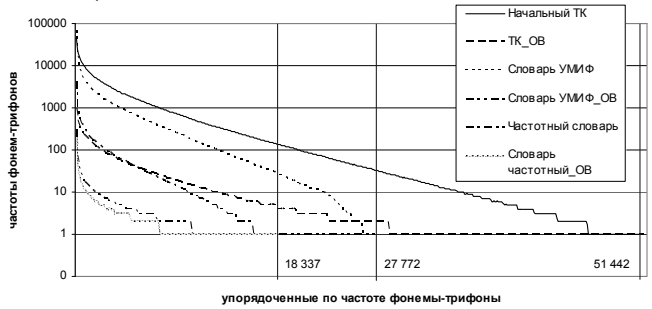

Рис. 2. Распределения фонем–трифонов на логарифмической шкале частотности в текстовых выборках

Для проверки эффективности использования речевых образов, т.е. фонем, открытых слогов и слогов, полученных по правилам деления слогов, формировались тексты *контрольной выборки* (КВ) слитной речи и проводилась процедура записи в условиях, аналогичных записи ОВ.

Первый способ выбора текста КВ основан на том, чтобы проверить распознавание часто употребляемых слов, предложений, фраз, т.е. сформировать КВ по частоте фонем–трифонов – «*частотную*» КВ. Полученная КВ содержит три с половиной часа записи и позволяет отследить ошибки при распознавании фонем во всех типичных контекстах. Объем словаря составляет 3 тыс. слов. Общее количество реализаций слов – приблизительно 9 тыс.

Второй способ состоит в формировании КВ случайным образом из тех же текстов, из которых выбирался текст ОВ, но с запретом выбора тех предложений, которые вошли в ОВ. Полученная «*случайная*» КВ содержит четыре с половиной часа записи и наиболее типична для выбранной предметной области, т.е. ошибка распознавания в ней будет иметь наиболее характерное значение выбранной предметной области. Объем словаря составляет 10 тыс. слов. Общее количество реализаций слов – приблизительно 23 тыс.

Таблица 1. Сравнение количества элементов (в тыс.) в источнике и полученной соответствующей ОВ после применения ЖА

Источник, из которого выбиралась ОВ	Общее количество предложений (слов для словарей) до работы ЖА	Общее количество предложений (слов) после работы ЖА	Общее количество реализаций фонем–трифонов до работы ЖА	Общее количество реализаций фонем–трифонов после работы ЖА	Всего элементов алфавита фонем–трифонов
ТК	816,0	18,0	41 179,8	1 020,3	51,4
Словарь УМИФ	1 874,7	13,7	23 734,3	120,1	27,7
Частотный словарь	137,6	8,2	1 488,0	71,0	18,3

Последняя КВ выбиралась из текстов, не использованных ни для формирования предыдущих КВ, ни для ОВ. Для этого из сайта украиноязычной Википедии случайным образом выбрано 100 Мб текстов – КВ «*Википедия*» (три часа записи). Объем словаря составляет более 7 тыс. слов. Общее количество реализаций слов – 16 тыс.

Компенсация несоответствия шкал акустической и лингвистической моделей

Как следует из выражения (1), на уровне базовых элементов (фонем) акустический декодер, в частности, пытается найти последовательность элементов $\mathbf{q}_{1:L} = q_1,...,q_L$, которые наиболее правдоподобно генерируют последовательность наблюдаемых векторов $\boldsymbol{Y}_{1:T} = \boldsymbol{y}_1,...,\boldsymbol{y}_L$, исходя из интегральной меры схожести:

$$\hat{\boldsymbol{q}} = \arg\max_{\mathbf{q}}\{\log p(\boldsymbol{Y} \mid \boldsymbol{q}) + \\ + (\alpha \log(P(\boldsymbol{q})) + \beta \mid \boldsymbol{q} \mid)\}, \quad (2)$$

где α и β – коэффициенты, компенсирующие несоответствие шкалы акустической и лингвистической, составляющих модели распознавания. Поэтому на первом этапе проводились эксперименты с целью эмпирически подобрать параметры α и β, рекомендуемый диапазон которых составляет 0 – 20 и 0 – (–20) соответственно [2, 12].

В экспериментальных исследованиях оценивались показатели фонемной ошибки (*PER – Phoneme Error Rate*):

$$\%PER = 100\% - \frac{H - I}{N}100\%$$

и фонемной некорректности (*PIR – Phoneme Incorrectness Rate*):

$$\%PIR = 100\% - \frac{H}{N}100\%,$$

где H – количество правильно распознанных элементов, I – количество ошибочно вставленных элементов, N – общее количество произнесенных элементов.

Убывание *PER* происходит главным образом путем уменьшения ошибочно вставленных элементов. Рост некорректности обусловлен уменьшением правильно распознанных элементов. Из рисунков следует, что наименьшая фонемная ошибка достигается при значениях параметров α = 5 и β = –5. Показатель корректности *PIR* дал возможность определить, что надежность возросла вследствие сокращения числа вставок.

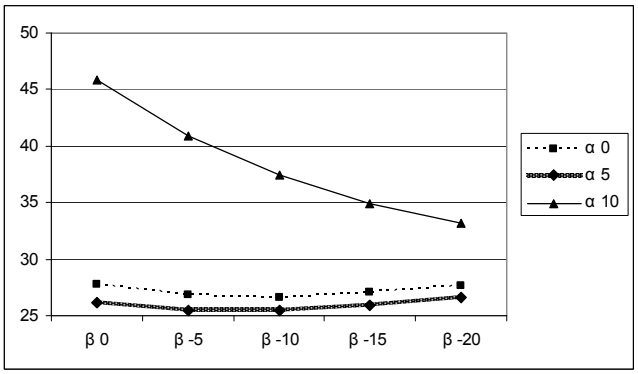

Рис. 3. Показатели *PER* распознавания (%) для слитной речи на «случайной» КВ

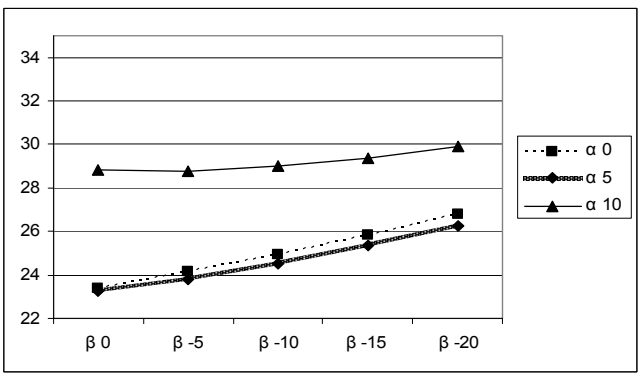

Рис. 4. Показатели *PIR* распознавания (%) для слитной речи на «случайной» КВ

Как следует из экспериментов, значимость рассмотренных параметров α и β не подлежит сомнению. Конечное решение об использовании тех или других значений зависит от того, что важнее: не потерять фонемы, которые потенциально могут совпадать с эталонами, или избавиться от как можно большего количества лишних элементов.

Увеличение эффективности акустической модели

В качестве композитного речевого элемента распознавания с помощью декодеров *HTK* и *Julius* [2, 12] были взяты фонемы (всего 59), открытые слоги (всего 7 270) и слоги, поде-

ленные по правилам украинского слогоделения (всего 10 200).

Были проведены эксперименты по исследованию влияния содержания ОВ акустической модели на распознавание. Рассматривались такие варианты акустической модели распознавания: модель, построенная только на слитной речи; модель, объединяющая слитную речь и изолированные слова; модель, которая не учитывает или учитывает ударение лишь частично.

При распознавании допускалась свободная грамматика следования фонемных образов как для фонем, так и для слогов. Ответы распознавания сводились к фонемному виду с целью дальнейшей оценки надежности в сравнении с эталонным фонемным текстом. В табл. 2 приведена фонемная ошибка *PER* для описанных ранее контрольных выборок. В результатах распознавания показана ошибка как с учетом, так и без учета фонемного ударения.

Таблица 2. Показатели ошибочно распознанных фонем (%) для слитной речи на основе различных речевых образов с использованием инструментария *HTK* и *Julius*.

Название КВ	Фонема		Открытый слог		Слог по правилам деления слогов	
	HTK	*Julius*	*HTK*	*Julius*	*HTK*	*Julius*
Случайная КВ	28,86	29,11	24,92	24,46	24,54	**24,03**
Случайная КВ (без ударения)	21,39	22,28	17,68	17,47	17,29	**17,01**
КВ Википедия	31,93	35,48	28,01	30,17	**28,18**	31,08
КВ Википедия (без ударения)	24,72	23,19	28,81	**20,81**	21,00	22,37
Частотная КВ	**36,6**	–	37,75	–	–	–
Частотная КВ (без ударения)	**26,1**	–	27,95	–	–	–

В табл. 3 представлены результаты фонемной ошибки распознавания при использовании различных акустических моделей. Сравнивая результаты, приведенные в таблице, видим, что использование акустической базы ОВ изолированных слов в дополнение к ОВ слитной речи приводит к уменьшению фонемной ошибки. Это можно объяснить тем, что акустическая база изолированных слов увеличивает количество реализаций каждой фонемы. Также наличие коротких синтагм способствует улучшению результатов распознавания, поскольку именно короткие синтагмы наиболее характерны для человеческой речи.

Акустические модели распознавания строились с учетом ударности гласных. На письме ударение обычно не указывается. В табл. 3 искусственным путем после распознавания были удалены все ударения. Это увеличило надежность распознавания на несколько процентов. А повлияет ли на надежность распознавания, если в акустической модели не учитывать ударение гласных? Для исследования этого была создана акустическая модель, которая игнорирует признак ударности в алфавите фонем.

Таблица 3. Показатели фонемной ошибки распознавания *PER* (%) для КВ слитной речи на основании разных речевых образов с использованием разных акустических моделей (слитную речь – СР и изолированные слова – ИС) инструментарием *HTK*.

Название КВ	Фонема		Открытый слог		Слог по правилам деления слогов	
	Используемая акустическая модель					
	СР	СР + ИС	СР	СР + ИС	СР	СР + ИС
Случайная КВ	28,86	25,6	24,92	23,06	24,54	**21,34**
Случайная КВ (без ударения)	21,39	20,96	17,68	**17,00**	17,29	18,36
Частотная КВ	36,6	26,7	37,75	23,22	–	**22,33**
Частотная КВ (без ударения)	26,1	21,56	27,95	**17,49**	–	18,11
КВ Википедия	31,93	30,76	**28,01**	28,53	28,18	29,35
КВ Википедия (без ударения)	24,72	24,52	28,81	22,02	**21,00**	22,88

Также, чтобы учесть специфику украинского произношения, а именно редуцирование безударных *е*, *и* к $е^{и}$, $и^{е}$ соответственно, была соз-

дана акустическая модель, в которой были оставлены только две ударные гласные *е́* и *и́*.

В табл. 4 приведены показатели ошибки распознавания *PER* (%) фонемного распознавания при использовании упомянутых акустических моделей фонем.

Таблица 4. Показатели фонемной ошибки распознавания *PER* (%) без учета ударности на различных акустических моделях

Название КВ	Используемая акустическая модель					
	СР	СР (ударение удалялось после распознавания)	СР + ИС	СР + ИС (ударение удалялось после распознавания)	СР + ИС (без ударения при обучении)	СР + ИС (ударные только *е* и *и*)
Случайная КВ	36,6	26,1	25,6	21,56	11,35	**11,27**
Частотная КВ	28,86	21,39	26,7	**20,96**	26,47	26,19
КВ Википедия	31,93	24,72	30,76	24,52	**21,43**	21,75

Из табл. 4 следует, что результаты зависят как от способа формирования акустической модели распознавания, так и от способа формирования КВ. Результаты, полученные на моделях только слитной речи, значительно улучшаются при объединении с моделями, построенными на изолированных словах. Следует обратить внимание на то, что результаты исследования (табл. 4), проводились только для фонемного распознавания, другие речевые образы не использовались.

Адаптация системы распознавания речи к различным платформам

В рамках Государственной научно-исследовательской программы «Образный компьютер» разработан ряд прототипов мобильных устройств, на которых реализованы технологии и алгоритмы распознавания и синтеза речевых сигналов. Вся линейка мобильных устройств (цифровой диктофон, голосовой секретарь и мобильный телефон) разработана на основе сигнальных процессоров *Analog Devices* семейства *BlackFin* [13]. Для этих процессоров существует возможность запуска на них операционной среды *uCLinux*, принадлежащей к семейству *UNIX*-подобных операционных систем и которая основывается на исходных кодах ядра ОС *Linux*. Используются три основных модуля – *GNU Toolchain* (кросс-компилятор), *Das U-boot* (исходные файлы загрузчика), *Linux Kernel* (исходные файлы ядра ОС *uCLinux*).

GNU Toolchain (кросс-компилятор) – специальный компилятор, работающий в операционной среде *Linux* на персональном компьютере и создающий исполняемый код для операционной среды *uCLinux* платформы на основе сигнального процессора *AD BlackFin*. Используется как для кросс-компиляции исходных кодов ядра ОС *uCLinux* и загрузчика *Das U-boot*, так и кросс-компиляции приложений, написанных на языке программирования *C\C++* для возможности их выполнения в среде *uC-Linux*. В состав компилятора входят следующие основные модули: компиляторы *gcc* и *gcc-elf* (версий 3.4 и 4.1, что дает широкие возможности совместимости программ) и специализированная библиотека для встраиваемых систем *uclibc*. Для удобства кросс-компилятор *GNU Toolchain* предоставляется в виде пакета *rpm* и в виде архивов *tar.gz*. Если система допускает работу с пакетами *rpm*, то удобнее использовать их. Также кросс-компилятор существует в двух версиях – для 32-битных и для 64-битных систем соответственно. После установки пакетов весь функционал будет доступен для использования стандартных методов процедуры *make*.

Das U-boot (загрузчик) – компьютерный загрузчик операционных систем, ориентированный на встроенные устройства архитектур *MIPS*, *ARM* и др. После кросс-компиляции может быть записан в *Flash-ROM* платформы. Затем код загрузчика выполняется при старте системы, позволяет загрузить в память и запустить ядро ОС *uCLinux*.

Linux Kernel (ядро ОС *uCLinux*) – центральная часть операционной среды *uCLinux*, обеспечивающая приложениям координированный доступ к ресурсам компьютера, таким как процессорное время, память и внешнее аппаратное обеспечение и реализирующая функции файловой системы.

Чтобы запустить на платформе, основанной на процессорах *BlackFin*, операционную среду *uCLinux*, необходимо сначала собрать конфигурацию загрузчика *U-boot* и ядра *Linux Kernel*, соответствующую функциональным и техническим характеристикам именно этой платформы, для чего следует вносить правки в конфигурационные файлы и (при необходимости) в файлы исходного кода загрузчика и ядра.

При распознавании отдельных слов система оперирует только словарем и акустическими моделями из базы данных и знаний. Распознавание слитной речи требует подключения порождающих грамматик в форме Бэкуса–Наура или статистической лингвистической модели [2]. В последнем случае декодер вначале использует биграммы, далее в узлах сформированного графа динамического программирования уточняются значения частичной меры схожести с привлечением N-грамм, $N > 2$.

Декодер реализован на языке программирования *C* [12] для персонального компьютера и адаптирован для возможности кросс-компиляции к микропрограммному коду операционной среды *uCLinux* сигнального процессора *BF*-561.

Результаты распознавания идентичных фрагментов речи на ПК и на портативных устройствах совпадают с точностью до шестого знака после запятой. Унификация программного кода позволяет все исследования проводить на персональном компьютере.

Разработана также система, реализующая клиент-серверную идеологию. При этом клиентские программы разрабатывались на языке *Java* для наиболее распространенной мобильной платформы *Android*. В клиентском ПО реализована возможность записи распознаваемого речевого сигнала, набора текста для дальнейшего озвучивания и обмена информацией со своими серверами. Серверное ПО разработано на языке *PHP* (получение данных от клиентов) и *C++* (распознавание и синтез речевых сигналов). Обмен данными между клиентом и сервером происходит по протоколу *http* с помощью стандартных процедур *POST* и *GET*. Объем словаря ограничивается вычислительными возможностями сервера.

Оба описанных подхода реализации распознавания речевых сигналов на мобильных устройствах – изолированный и клиент-серверный – закладывают предпосылки к введению гибридного подхода, в котором предусматривается попытка распознать речь непосредственно на мобильном устройстве, а в случае отказа распознавания – воспользоваться связью с сервером.

Заключение. Описанная модель пофонемного распознавания позволяет не только включать в словарь новые слова без обучения на них, но и особо актуальна при реализации схем распознавания с введением независимых уровней для языков с большим количеством словоформ и относительно свободным порядком следования слов, к которым относятся славянские языки.

Применение разработанной в ходе исследований компенсации несоответствия шкал акустической и лингвистической моделей, а также учет явления ударности гласных повышают эффективность системы распознавания. Предложенный способ формирования ОВ дает возможность широко охватить фонетическое разнообразие языка, используя менее двух процентов предложений из всех рассмотренных. В то же время генерирование многовариантных транскрипций описывает вариативность произношения каждого диктора.

Исследования показали, что при адаптации декодера к архитектуре портативных устройств наиболее гибкая – гибридная архитектура, позволяющая воспользоваться одновременно преимуществами изолированного и клиент-серверного подходов.

Дальнейшие исследования следует посвятить моделированию взаимовлияния звуков в потоке речи, индивидуализации параметров модели распознавания и интеграции с технологией озвучивания текстов.

1. *Винцюк Т.К.* Анализ, распознавание и смысловая интерпретация речевых сигналов. – Киев: Наук. думка, 1987. – 263 с.

2. *Gales M*., *Young S*. The Application of Hidden Markov Models in Speech Recognition // Foundations and Trends in Signal Processing. – 2007. – N 1(3). – P. 195–304.
3. *Vintsiuk T*., *Sazhok M*. Multi-Level Multi-Decision Models in ASR. // Proc. of the 10th Int. Workshop «Speech and Computer» (SPECOM'2005). – Patras, 2005. – P. 69–76.
4. *http*://www.creaceed.com/ceedvocal/about
5. *http*://www.apple.com/ios/siri/
6. *http*://www.nuance.com/for-business/by-product/automotive-products-services/vocon-hybrid/
7. *Робейко В.В*., *Сажок М.М*. Багатозначна багаторівнева модель перетворення орфографічного тексту на фонемний // Штучний інтелект. – 2011. – № 4. – С. 117–125.
8. *Широков В.А*., *Манако В.В*. Організація ресурсів національної словникової бази. // Мовознавство. – 2001. – № 5. – С. 3–13.
9. *Робейко В.В*., *Сажок М.М.* Використання текстового корпусу для прогнозування наголосів у словах української мови. // Інтелектуальні системи прийняття рішень та проблеми обчислювального інтелекту: Матеріали міжнар. наук. конф. – Херсон, 2012. – С. 171–172.
10. *Сажок Н.Н*. Кластеризация слов при построении лингвистической модели для автоматического распознавания речевого сигнала. // Кибернетика и вычислительная техника. – 2012. – **170**. – С. 59–66.
11. *Goncharov E*., *Kochetov Yu*. Behavior of probabilistic greedy algorithm for stage location problem // Sampling analysis and operations research. – 1999. – **6**, N 1. – P. 12–32.
12. *Lee A*., *Kawahara T*. Julius – an open source real-time large vocabulary recognition engine. // Proc. Europ. Conf. on Speech Communication and Technology (EUROSPEECH) – 2001. – P. 1691–1694.
13. *Розроблення* програмно-апаратних засобів базового модуля усномовної комп'ютерної технології, що вбудовується в сучасні комп'ютерні системи, створення на їх основі високотехнологічних електронних виробів широкого застосування та здійснення заходів для їх впровадження в виробництво (ОК_2009_2): Звіт про НДР МННЦІТіС НАН та МОН України. – Київ, 2010. – 149 с. – № ДР 0109U004244.

Тел. для справок: +38 (044) 502-6333 (Киев)
E-mail: *n.vassilleva@gmail.com*, *valya.robeiko@gmail.com*, *dmytro.fedoryn@gmail.com*