УДК 004.934

Н.Н. Сажок, В.В. Робейко, Д.Я. Федорин, Р.А. Селюх

Система преобразования телерадиовещания в текст для украинского языка

Описаны система превращения сигнала телерадиовещания в текст для украинского языка и моделирование особенностей, специфических для него – нерегулярность лексического ударения и высокая флективность. Разработанная система реализует подход клиент–сервер и позволяет просматривать пятиминутные сегменты речи синхронно с результатом распознавания речи.

Ключевые слова: распознавание речи, телерадиовещание, языковые особенности, украинский язык.

Описано систему перетворення сигналу телерадіомовлення в текст для української мови та моделювання особливостей, специфічних для неї – нерегулярність лексичного наголосу та висока флективність. Розроблена система реалізує підхід клієнт–сервер і дає змогу переглядати п'ятихвилинні сегменти мовлення синхронно з результатом розпізнавання мови.

Ключові слова: розпізнавання мовлення, телерадіомовлення, мовні особливості, українська мова.

Введение. Распознавание речи находит новые сферы применения в информационном обществе. Одна из таких сфер – обработка медийной информации, в частности теле- и радиовещания. Существует ряд преимущественно экспериментальных и вспомогательных систем, в которых автоматизированы генерирование субтитров и поиск информации для английского и других языков, в основном европейских [1, 2]. В основе таких систем лежит технология распознавания речи, предназначенная для преобразования в текст сигнала, принимаемого из определенных источников вещания и соответствующая определенному набору телерадиопрограмм (новости, интервью, телешоу, трансляция заседаний парламента и пр.). Полученный в результате преобразования текст должен соответствовать содержанию, а пользователь системы должен иметь возможность прослушивать запись передачи, параллельно следя за текстом и по мере необходимости корректируя его. При этом важно сократить задержку получения ответа распознавания, одновременно учитывая ограничения доступных вычислительных ресурсов. Сегодня не существует системы преобразования украинского телерадиовещания в текст для последующего его анализа.

Наиболее специфичны для украинского, как и для любого другого славянского языка, высокая флективность и относительно свободный порядок слов, что приводит к быстрому росту словаря распознавания (в восемь–10 раз больше, чем для английского языка в такой же предметной области) и ослаблению предиктивной силы при моделировании допустимого следования слов. Поэтому прямое применение общепринятых методов и алгоритмов к славянским языкам не является многообещающим, что в свою очередь стимулирует разработку альтернативных схем, основанных, в частности, на композиции слов по результатам пофонемного декодирования [3]. Вместе с тем, потенциал апробированной десятилетиями исследований схемы все еще остается до конца нераскрытым [4]. Так, не исследованы ограничения на объем словаря, который используется в системе преобразования речи в текст на основе общепринятой схемы, исходя из того, что система должна демонстрировать продуктивность, сопоставимую с реальным временем на современных вычислительных платформах.

Таким образом, авторы задались целью разработать систему, способную оперативно преобразовывать в текст украинскую речь, записываемую из некоторого множества каналов телерадиовещания с последующей возможностью просмотра и редактирования результата распознавания через Интернет. Словарь системы должен покрывать произвольный текст из расчета менее определенного процента внесловарных слов (*OOV*), должны быть предоставлены средства пополнения словаря. Не рассматриваются как показательные результаты распознавания

при искажениях акустического сигнала, вызванных некачественной записью и/или существенной потерей данных вследствие сжатия сигнала, а также в сегментах речи со значительными шумовыми помехами и наложением нескольких источников речи.

В предыдущих работах авторы исследовали особенности распознавания украинской спонтанной речи в реальном времени, описывали системы преобразования речи в текст для общих и новостных предметных областей, относящихся к политике, экономике, культуре и на ряде вычислительных платформ [5, 6].

В данной статье авторы объясняют допущения, касающиеся языковых особенностей на акустическом, фонетическом и лексическом уровнях, прокладывают пути к достижению необходимого объема словаря, описывают соответственный разработанный программный инструментарий вместе с экспериментальными исследованиями, а также систему преобразования телерадиовещания в текст в целом.

Общая структура системы преобразования текста в речь

Структура системы показана на рис. 1. Компонента реального времени *Распознаватель* обращается к *Базе данных* (Д) *и знаний* (З), формируемой офф-лайн с помощью средств, не вошедших в иллюстрацию. Для создания указанных в структуре компонент авторами разработан ряд программных ресурсов и ресурсов данных, а также использован разного рода инструментарий, доступный в Интернете.

Компонента реального времени получает *Входящий речевой сигнал* из некоторого источника (в данном случае – сеть *IPTV* или файловая система). *Детектор голосовой активности* обнаруживает предполагаемые начала речевых сегментов, чтобы начать передачу сигнала в *Препроцессор*, извлекающий первичные акустические признаки. При этом используются мел-кепстральные коэффициенты с вычтенным средним и дополненные энергией и динамическими компонентами (Δ и ΔΔ-коэффициенты). *Декодер* сравнивает входящий сегмент с гипотезами модельного сигнала, которые генерируются на основе акустической и лингвистической моделей с использованием консервативной стратегии отбрасывания неперспективных гипотез [7]. Результат декодирования, представленный в виде последовательностей слов или сети несовпадений, дополненных оценками длительностей и доверительной мерой, передается в *Блок принятия решений* (БПР), формирующий окончательный Ответ распознавания с учетом предыстории и доверительных интервалов.

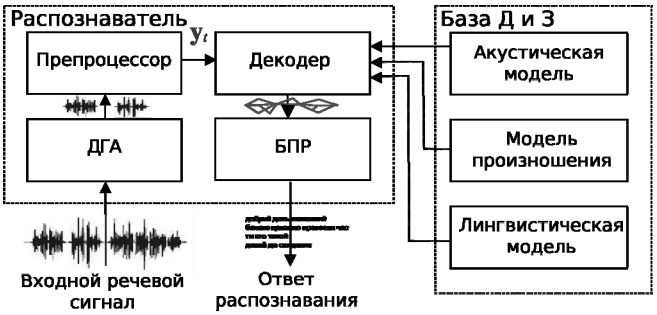

Рис. 1. Общая структура базовой системы преобразования текста в речь

Акустическая модель разработана на 40-часовом материале корпуса АКУЕМ [8, 9]. Базовый алфавит фонем насчитывает 56 фонем, включая ударные и безударные версии для шести гласных. Необходимость их различать обосновывается далее. Области пребывания фонем в первичном пространстве признаков описываются гауссоидами (от восьми до 32) в соответствующих эталонах.

Модель произношения предоставляет *Декодеру* фонемные транскрипции слов, сформированные офф-лайн модулем графемно-фонемного преобразования, в котором реализован метод многозначного преобразования символов, основанный на задании отношений между орфографическими и фонемными символами [10]. Эксперту достаточно сформулировать около 40 локальных правил перехода от графем к фонемам, в которых определенным образом отображены как индивидуальные особенности произношения, так и коартикуляция и редукция звуков в потоке речи. Правила настроены так, что в среднем для каждого слова генерируется 1,2 транскрипций. Такой же алгоритм, но с другими правилами, применяется для преобразования чисел, сокращений и неалфавитных графем в последовательности слов. Словарь для

системы состоит из частотного словаря, извлеченного из текстового корпуса, и дополнительных подсловарей, покрывающих речевой корпус, социальные и местные диалекты, имена собственные, аббревиатуры и пр. Рабочий словарь распознавания формируется путем отбора наиболее частотных слов из словаря с учетом предметной области.

Лингвистическая модель создана исходя из словаря распознавания и подмножества корпуса текстов, состоящего из предложений, содержащих ниже определенной части *OOV*-слов. Основной текстовый корпус является производным от гипертекстовых данных, загруженных из нескольких вебсайтов, содержащих образцы новостей и рекламы (60 процентов), литературы (8 процентов), энциклопедических статей (24 процента), в правовой и судебной области (8 процентов). Следует отметить, что данные, загруженные из новостных сайтов, содержат многочисленные комментарии пользователей и отзывы, которые рассматриваются как текстовые образцы спонтанной речи. Текстовый фильтр, используемый для обработки корпуса текстов, обеспечивает преобразование чисел и символьных графем в последовательности букв, удаляя неправильные сегменты текста, и повторяющихся абзацев. Общий размер базового текстового корпуса составляет 2 Гб, что включает в себя 17,5 млн предложений, что соответствует списку слов, содержащих более 275 млн единиц, и образующих словарь более двух миллионов слов.

Для словаря распознавания на 100 тыс. слов, зафиксировано 88,5 млн различных трехграмм в подкорпусе основного текстового корпуса после удаления предложений, содержащих более 20 процентов или по крайней мере три последовательных неизвестных слова. Этот подкорпус, используемый для моделирования допустимых последовательностей слов, будем обозначать 250M. Примечательно, что в 250M *OOV*-слова занимают 2,5 процента всех слов, что примерно в два раза меньше, чем в украинском произвольном тексте для указанного размера словаря. Для моделирования спонтанных характеристик речи в словарь распознавания введен класс прозрачных слов, содержащих нелексические единицы, такие как заполненная пауза и выражения эмоций и отношения (смех, аплодисменты и др.).

Применяя инструментарий лингвистического моделирования [11], авторами получен текстовый файл в формате *ARPA* размером 5 Гб, который был уменьшен до 1,2 Гб вследствие применения модуля бинаризации из инструментария декодера [7].

Модули реального времени использованы для построения базовой системы преобразования речи в текст с целью проведения экспериментальных исследований и опытной эксплуатации. Интеграция базовой системы с графическим интерфейсом позволила демонстрировать систему диктовки слитной речью для широкой предметной области в реальном времени на современном ноутбуке [6].

Рассмотрим признаки, специфические для украинского языка, чтобы обосновать предположения относительно моделирования языковых особенностей на акустическом, фонетическом и лексическом уровнях и расширения базовой системы преобразования из речи в текст.

Анализ лексического ударения

Во многих языках определенные слоги в словах более ярко выражены в просодических терминах, таких как длительность, основной тон и громкость звука. Это явление называется *лексическим ударением*. Следует ли вводить отдельно ударные и безударные гласные в базовый алфавит фонем?

В отличие от ряда европейских языков, давая позитивный ответ на этот вопрос, будем полагаться на фонетические, лексические и акустические знания об украинском языке. Ударность в гласных обычно действует как и изменения в фонемном составе слов: слово может приобрести иную грамматическую форму или смысл, что прослеживается примерно в 10 процентах слов произвольного текста.

Для исследования акустической стороны вопроса была проведена оценка параметров акустических моделей ударных и безударных гласных так, как если бы это были разные фонемы. Далее были изучены различия между моделями, в частности, средствами визуализации гене-

ративных моделей HMM [12]. На рис. 2 разница между ударной и безударной фонемами *а* наблюдается в определенных составляющих моделей. Визуализация моделей других фонем доступна на веб-странице упомянутого инструментария.

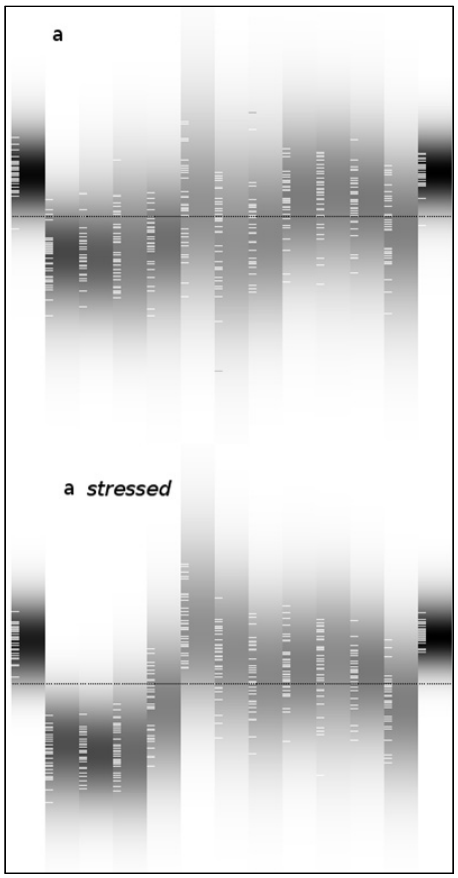

Рис. 2. Визуализация акустических моделей украинских безударной и ударной фонем **а**

В украинском языке позиция ударения нерегулярна и может изменяться даже среди форм одного и того же слова. Очевидно, экспертное указание положения ударения для всего лексикона очень трудоемко, а потому неприемлемо. Поэтому авторами предложена процедура предсказания ударной позиции в слове на основе известного словаря ударений и текстового корпуса.

Рассматриваются все допустимые сегментации S для слова с неизвестным ударением. Сегментация i-я

$$S_i = \left(q_{i,1}, q_{i,2}, ..., q_{i,j}, ..., q_{i,L_i} \right) \qquad (1)$$

имеет длину L_i. Здесь $q_{i,j}$ является j-м элементом (буквой или фонемой) в i-й сегментации. Далее введем вектор θ_{L_i} – индикатор уровня ударности (например, ноль, один, два) для каждого из L_i элементов. Теперь можно оценить вероятность ударной позиции при условии сегментации S_i:

$$P\left(\theta_{L_i} \mid S_i\right) \approx \frac{c\left(S_i, \theta_{L_i}\right)}{c\left(S_i\right)}, \qquad (2)$$

где $c\left(S_i, \theta_{L_i}\right)$ – количество сегментов в S_i, имеющих ударение, определенное вектором–индикатором θ_{L_i}, а $c(S_i)$ – общее число наблюдений S_i. Все подсчеты проводятся по текстовому корпусу за исключением слов, не вошедших в словарь ударений.

Наконец, проводим поиск по всем допустимым сегментациям S и положениям ударения θ^S, которые доставляют

$$\arg\max_{S,\theta^S} \prod_{S_i,\theta_{L_i}} P\left(\theta_{L_i} \mid S_i\right). \qquad (3)$$

Итак, сконструирован граф динамического программирования, в котором нахождение кратчайшего пути соответствует поиску (3). Запоминая N перспективных стрелок, входящих в узлы этого графа, можно находить N лучших ударных положений, дополняемых оценкой вероятности.

Параметры модели предсказания ударений на текстовом корпусе 250 М получили оценку. Дополнительно введен символ границы между словами. Обнаружено более 60 тыс. символьных сегментов длины от одного до четырех. На рис. 3 приведен пример однозначного прогнозирования ударения для имени собственного *Обама*, которое отсутствует в базовом украинском словаре. Слово представлено как конкатенация всех допустимых символьных сегментов, где наибольшая длина сегмента ограничена четырьмя символами. Каждый входной символ вводит множество допустимых сегментов. На рис. 3 потенциально оптимальные частичные траектории полностью показаны для колонок 1 и 2, в других колонках указано имя узла, из которого исходит потенциально оптимальная траектория. Указанные частичные критерии основаны на логарифме вероятности. Оптимальная

траектория |*o-b-Ama*|, соответствующие узлы и критерии выделены жирным. Следует обратить внимание на то, что в колонке 7 потенциально оптимальная траектория в узел *Ama*| входит из узла *a*, а не из *obA* – узла с лучшим критерием, чем *a*. Узел *obA* отброшен для того, чтоб избежать двух подряд идущих ударений в слове.

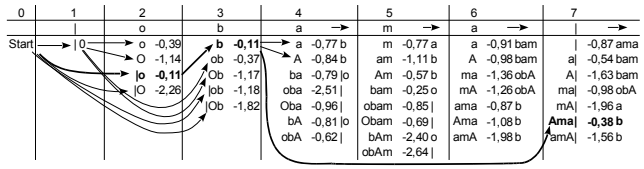

Рис. 3. Прогнозирование ударения для отсутствующего в словаре ударений слова *obama*

Процедура оценки ошибки предсказания ударения не так очевидна, как казалось бы, поскольку не всегда однозначен ответ на вопрос, что есть ошибка в ряде конкретных случаев. Например, считать ли ошибкой, если ударение предсказано не точно в словах с опечаткой? В любом случае, предварительные исследования показали уровень ошибки между 5 и 10 процентами относительно размера словаря.

Разработка лингвистической модели на основе классов слов

Как и любой другой славянский язык, украинский является флективным, что приводит к наличию для каждого слова в среднем 12 словоформ, что в шесть раз больше, чем в английском языке. Поэтому для построения лингвистической модели, охватывающей сопоставимый лексикон, для украинского языка требуется соответственно словарь с объемом, большим в шесть раз. Более того, относительно свободный порядок следования слов приводит к росту разветвленности и разреженности. Анализ этих свойств мотивировал к переходу от статистической лингвистической модели, оперирующей словами, к модели, оперирующей классами эквивалентности и вероятностью принадлежности слова к классу [13].

При соотнесении слов к классам (кластеризации), предпринимается попытка минимизировать критерий разветвленности:

$$F_G = \sum_{g,h \in G} C(g,h)\log C(g,h) - 2\sum_{g \in G} C(g)\log C(g), \quad (4)$$

где (g,h) означает, что класс g следует за классом h из множества классов эквивалентности G, а функция $C(\cdot)$ вычисляет частоту наблюдения аргумента в обучающей выборке. В алгоритме обмена [13] предполагается множество итераций, где для каждого слова тестируется его принадлежность ко всем классам с последующим соотнесением к тому классу, для которого достигнут наилучший критерий (4). В процессе реализации этого алгоритма была предложена альтернативная формулировка ускорения вычисления критерия (4) [14].

Результаты кластеризации анализировались на соответствие лингвистическим категориям. Впервые полученные автоматическим путем классы для украинского языка продемонстрировали в подавляющем большинстве случаев однородность по синтаксическим, семантическим и, в меньшей мере, фонетическим признакам.

Большинство классов имеют очевидную синтаксическую интерпретацию. Так, одни классы могут содержать имена существительные родительного падежа, другие – прилагательные множественного числа и пр.

Несколько классов слов, полученных в результате биграммной кластеризации на корпусе 250 М для 1000 классов показаны в табл. 1.

Таблица 1. Пример биграммной кластеризации, $G = 1000$

Слово в кластере с переводом	Частота
Багато / много	134590
Чимало / немало	24482
Безліч / множество [чего-л.]	7696
Немало / немало	2191
Якнайбільше / как можно больше	760
Багацько /множество [чего-л.]	255
Богато (*ош.* багато)	123
Які / какие	590681
Котрі / которые	24499
Яки (*ош.* які)	465
Де / где	246376
Куди / куда	31966
Звідки / откуда	15373
Звідкіль / откудова	120
Заявив / заявил	163547
Вважає / полагает	99803
Повідомив / информировал	80043
Заявила / заявила	32795
Заявляє / заявляет	31965
Розповів / рассказал	30504
Говорить / говорит	29756

Слова в каждом классе отсортировованы по убыванию их частотности, а наиболее встре-

чаемое слово в классе выделено полужирным шрифтом. Полностью представлены три класса, а для последнего в таблице класса приведены лишь первые семь слов. Этот класс служит примером семантической однородности: в нем содержатся глаголы третьего лица, имеющие смысл коммуникации. Два первых класса демонстрируют, что ошибочно введенные, но, тем не менее, достаточно частотные слова (*богато* и *яки*) отнесены к классам, в которые вошли слова с правильным написанием.

Результаты кластеризации также показали определенную чувствительность к фонетическому наполнению слов. Например, в украинском языке союз *и* передается одной из трех форм в зависимости от фонетического окружения: между гласными, между согласными и в других случаях. И каждая из этих трех форм была автоматически отнесена к различным классам.

В *экспериментальных исследованиях* рассмотрены две *контрольные выборки* (КВ). КВ 1 содержит 49 предварительно отобранных тематически сбалансированных записей, в КВ 2 вошло 78 случайных записей. Из табл. 2 следует, что обе выборки по длине примерно одинаковы, а КВ 1 более ориентирована на судебную предметную область. Для каждой контрольной выборки оценены акустические параметры на реализациях корпуса АКУЕМ, не вошедших в соответствующую контрольную выборку.

Таблица 2. Характеристика контрольных выборок

КВ	Длина (час)	Судебные шоу, %	Речь судьи, %	Новости, %	Токшоу, %	Пресс-конференции,%
1	11,4	69,4	11,1	8,4	8,2	2,9
2	12,6	32,5	–	29,8	36,8	0,90

Лингвистическая модель построена на текстовом корпусе 250 М. Не использованы предложения из транскрипций текстовых АКУЕМ. По наиболее частотным словам сформированы словари 100*k* и 200*k*, содержащие соответственно 100 и 200 тыс. слов. Только первые по частотности 100 тыс. слов прошли кластеризацию. Менее частотные слова были соотнесены к классу «неизвестных» слов.

Как следует из табл. 3, использование ударных гласных ведет к заметному уменьшению ошибок (*WER* – показатель пословных ошибок). Несмотря на незначительное снижение, в сравнении с моделями, основанными на словах, лингвистическая модель на основе классов демонстрирует определенный потенциал, который заключается в уменьшении требований к оперативной памяти и в лучших перспективах – по увеличению объема словаря.

Таблица 3. Экспериментальные результаты

КВ	Ударение	Классы	Порядок ЛМ	Размер словаря / %*OOV*	%*WER*
1	–	–	3	100*k* / 5,27	33,6
1	+	–	3	100*k* / 5,27	32,1
1	+	1000	3	200*k* / 3,79	34,1
1	+	1000	4	200*k* / 3,79	33,8
2	–	–	3	100*k* / 5,61	38,0
2	+	–	3	100*k* / 5,61	36,3
2	+	1000	3	200*k* / 4,15	38,7
2	+	1000	4	200*k* / 4,15	38,5

Опытная эксплуатация системы преобразования телерадиовещания в текст

Создание систем автоматической обработки речи – одно из наиболее актуальных направлений развития современных информационных технологий. В зависимости от места, где происходит преобразование *произнесенная фраза – текст* и *текст – произнесенная фраза*, системы автоматической обработки речи делятся на изолированные (*client-side*), клиент–серверные (*server–side*) и гибридные (*hybrid*). В изолированных системах все преобразования происходят непосредственно на клиентском устройстве. В клиент–серверных – клиентское устройство используется только для ввода информации, передачи ее по сети на сервер для дальнейшей обработки и получения от сервера ответа распознавания. Гибридные системы совмещают в себе функционал изолированных и клиент–серверных – при наличии доступа к сети они используют для преобразования сервер, при недоступности сети работают как изолированная система.

Каждый из подходов имеет свои преимущества и недостатки. Изолированная система ограничена быстродействием и размером доступной оперативной памяти современных мо-

бильных систем, что в свою очередь накладывает ограничения на размер словаря и увеличивает время ответа приложения. Клиент–серверная технология не имеет этих ограничений, но требует для работы постоянного подключения к глобальной сети. Гибридная технология – это, по сути, реализация двух предыдущих технологий в одной системе, поэтому ее разработка требует больше времени и ресурсов, чем реализация каждой из технологий отдельно.

Примером реализации изолированной (*client-side*) системы распознавания служит линейка мобильных устройств – цифровой диктофон, голосовой секретарь и мобильный телефон. Эти устройства разрабатывались в рамках Государственной научно-исследовательской программы «Образный компьютер» [15] на базе сигнальных процессоров *Analog Devices* семейства *BlackFin*. Быстродействие процессора *BlackFin* и недостаточный объем оперативной памяти портативных устройств не позволяли увеличить словарь для распознавания речи выше 10–15 тыс. слов без существенного замедления процесса получения ответа распознавания.

Для распознавания произвольной слитной речи нужны гораздо большие объемы лексикона. Развитие системы распознавания [5] в клиент–серверном направлении, дало возможность в полной мере использовать описанные ранее теоретические наработки и, таким образом, перейти к словарям, содержащим сотни тысяч слов, в целом расширяя сферу применения технологии преобразования речи в текст.

Разработанная система предусматривает обмен данными между клиентом и сервером распознавания через сеть (Интернет либо локальную) по протоколу *TCP*/*IP*. Обмен происходит с использованием *REST*-интерфейса, т.е. вызов удаленной процедуры представляет собой обычный *HTTP*-запрос (*POST* или *GET*), а необходимые данные передаются в качестве параметров запроса. Серверное программное обеспечение разрабатывалось на языках *C++* (распознавание речевых сигналов), *PERL* (взаимодействие с аудиовидеоданными) и *PHP* (обработка запросов от клиентов). В интерфейсе клиента используются возможности *Java-Script* и *HTML5*.

Сегодня в сети Интернет на основе разработанной системы доступен экспериментальный интерфейс преобразования украинского телерадиовещания в текст [16]. Пользователь выбирает один из нескольких каналов и получает вебстраницу с результатом преобразования последнего записанного пятиминутного фрагмента в текст с обеспечением синхронизации с медиапроигрывателем, т.е. при проигрывании медиафайла синхронно подсвечивается именно то слово из распознанного текста, которое, по мнению системы, соответствует текущему сегменту речи.

Выбор пользователем слова *залишилися* показан на рис. 4. При этом проигрывается сегмент речи, соответствующий выбранному слову. Прослушав сегмент, пользователь исправляет результат распознавания, внося правку в окне редактирования рядом с выбранным словом. Слева от текста под медиа-проигрывателем размещены элементы управления, позволяющие передвигаться как между соседними фрагментами, так и с интервалом, равным одному часу и суткам.

Рис. 4. Визуализация результата преобразования фрагмента телепередачи в текст

Объем словаря системы составляет более 200 тыс. слов. Распознавание происходит вдвое быстрее реального времени, т.е. запись продолжительностью в одну минуту распознается примерно за 30 с. Система использует мощности четырехядерного процессора *Intel Xeon*, поэтому возможно одновременное выполнение восьми различных задач распознавания без потери быстродействия.

Заключение. Разработанная технология преобразования телерадиовещания в текст демонстрирует потенциал использования отличительных особенностей языка для приближения к моделированию всего объема лексикона с перспективой автоматической расстановки пунктуации и определения регистра символов в результате распознавания. Предложенная процедура расстановки лексических ударений не только способствует улучшению результатов распознавания, но и слу-

жит развитию исследований морфемного и семантического уровней в распознавании речи. Использование классов слов в лингвистической модели существенно уменьшает потребности в ресурсах оперативной памяти и упрощает процесс пополнения рабочего словаря системы распознавания новыми словами. Созданная экспериментальная система преобразования телерадиовещания в текст открывает путь к автоматическому анализу информационного потока телерадиовещания в Украине. Дальнейшее развитие системы предполагает введение многоязычности и добавление новых функций, таких как поиск по ключевым словам, получение метаданных, сопровождение дикторов (*speaker diarization*), разбивка на тематические сюжеты и др. Актуальными остаются вопросы улучшения надежности распознавания, особенно при получении акустического сигнала с искажениями и шумами.

1. *http*://voxalead.labs.exalead.com/
2. *http*://tech.ebu.ch/docs/events/metadata15/ Petr Vitek and Pavel Ircing_CT_UWB.pdf
3. *Vintsiuk T.*, *Sazhok N.* Multi-Level Multi-Decision Models for ASR // Proc. SpeCom'2005. – Patras, 2005. – P. 69–76.
4. *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.
5. *Sazhok N.*, *Robeiko V.*, *Fedoryn D.* Distinctive features for Ukrainian real-time speech recognition system // Proc. UkrObraz'2014. – Kyiv, 2014. – P. 66–70.
6. *Robeiko V.*, *Sazhok N.* Real-time spontaneous Ukrainian speech recognition system based on word acoustic composite models // Proc. UkrObraz'2012. – Kyiv, 2012. – P. 77–81.
7. *Lee A.*, *Kawahara T.* Recent Development of Open-Source Speech Recognition Engine Julius. APSIPA ASC, 2009. – P. 131–137.
8. *The HTK* Book Version 3.4 / S. Young, G. Everman, M. Gale et al. – Cambridge University, 2006. – 359 p.
9. *Ukrainian* Broadcast Speech Corpus Development / V. Pylypenko, V. Robeiko, N. Sazhok et al. // Specom'2011. – Kazan. – P. 244–247.
10. *Robeiko V.*, *Sazhok N.* Bidirectional Text-To-Pronunciation Conversion with Word Stress Prediction for Ukrainian // Proc. UkrObraz'2012. – Kyiv, 2012. – P. 43–46.
11. *Bo-June* (*Paul*) *Hsu*, *James Glass*. Iterative Language Model Estimation: Efficient Data Structure & Algorithms // Proc. Interspeech, 2008.
12. *www*.cybermova.com/speech/visual-hmm.htm
13. *Martin S.*, *Liermann J.*, *Ney H.* Algorithms for bigram and trigram word clustering // Proc. of Eurospeech. – Madrid, 1995. – **2**. – P. 1253–1256.
14. *Сажок Н.* Кластеризация слов при построении лингвистической модели для автоматического распознавания речевого сигнала // Киберентика и вычислительная техника. – 2012. – № 4. – С. 59–66.
15. *http*://obrazcomp.irtc.org.ua/Osn_resultat.html
16. *www*.cybermova.com/technology/synchrophone.html

E-mail: *sazhok@gmail.com*, *valya.robeiko@gmail.com*, *dmytro.fedoryn@gmail.com*, *vxml*12*@gmail.com*

UDC 004.934

N.N. Sazhok, V.V. Robeiko, D.Ya. Fedoryn, R.A. Selyukh

Broadcast Speech-to-Text System for the Ukrainian

Keywords: speech recognition, broadcast, language-specific features, Ukrainian.

Introduction: Broadcast data processing is an important task for information society. The experience in development of real-time systems for Ukrainian dictation and speech record recognition on several computational platforms is the base for the described R&D devoted to extracting text from broadcast speech signal.

Methods: The modeling is focused on features that are specific particularly for Ukrainian such as lexical stress and high inflexibility. Given arguments confirm the necessity to distinguish stressed and unstressed vowels in the phoneme alphabet. Lexical stress irregularity implies expert involvement for stress assignment. To automate this procedure we implemented a data-driven stress prediction algorithm that represents words as sequences of substrings and searches for one or more sequences with the best criteria. As a Slavonic language Ukrainian is highly inflective and tolerates relatively free word order, which motivates transition from word- to class-based statistical language model.

Experimental research: Modeling both stressed and unstressed vowels leads to recognition accuracy improvement. Introduction word equivalence classes to the Language Model significantly decreases RAM consumption keeping the same recognition accuracy level. The developed experimental system implements client–server approach and allows for browsing 5-minute broadcast segments synchronously with speech recognition result.

Conclusion: Language-specific speech feature modeling is beneficial for a speech recognition system. The created broadcast speech-to-text system opens news prospectives for broadcast stream analysis in Ukraine.

●