Корпус украинской эфирной речи

Васильева Н.Б., *научный сотрудник*
Пилипенко В.В., *старший научный сотрудник*
Радуцкий А.М., *кандидат технических наук*
Робейко В.В., *научный сотрудник*
Сажок Н.Н., *кандидат технических наук, старший научный сотрудник*

В статье описывается основанный на эфирных записях корпус украинской разговорной речи. Информация, содержащаяся в корпусе, используется для создания акустической и языковой модели в технологии распознавания речи. Изложен краткий обзор текущего состояния разработок речевых корпусов в мире. Даны характеристики созданных в Украине речевых корпусов, имеющих практическое или историческое значение. Обсуждается роль и особенности использования текстовых и речевых корпусов в современных речевых технологиях, указываются наиболее характерные ошибки и противоречия, возникающие при формулировании концепции корпуса. Представлены концепция большого репрезентативного корпуса разговорной речи и требования к его свойствам. Описаны фундаментальные понятия и технические решения, используемые при записи и аннотировании речевого материала. Введенная система обозначений обеспечивает подробное описание зарегистрированных данных. Определены первоочередные задачи, стоящие на пути расширения корпуса. Экспериментальные исследования текущей версии речевого корпуса показали устойчивое повышение надежности распознавания речи по сравнению с более ранними стадиями разработки корпуса. Представленный материал рассматривается как шаг к созданию национального речевого корпуса, применимого для разработки широкого диапазона речевых технологий.

• *корпус спонтанной речи* • *распознавание речи* • *сегментирование* • *аннотирование*.

In this paper we describe a media-based speech corpus for spoken Ukrainian language. Information contained in the corpus is aimed to develop acoustic and language models for speech recognition technology. We give an overview of the current state of the art in speech corpora all over the world. Developed in Ukraine speech databases both historical and available today are listed and summarized. Nowadays role and specific features of text and speech corpora are investigated as well as the most frequent mistakes and misunderstandings of the corpus concept are discussed. The concept of a large representative corpus of spoken language and its desired properties are presented. Basic concepts and technical solutions used for speech corpus recording and annotation are described. The introduced mark-up system provides a detailed description of the recorded data. The most significant problems standing in the way of building a huge speech corpus are pointed out. A current version of the speech corpus has been validated with HTK tools that showed steady progress of speech recognition accuracy comparing to early stages of corpus development. We consider the presented corpus as a step to creation the national speech corpus applicable for entire range of speech technology.

• *spontaneous speech* • *speech recognition* • *segmentation* • *annotation*.

Вступление

Речевые корпуса играют большую роль при разработке речевых информационных технологий. Информация, которая содержится в таких корпусах, используется для построения акустических и лингвистических моделей для построения как систем наговаривания речи, так и моделей диалога человека с машиной, а также моделей предметных областей для смысловой интерпретации речи. Особые требования предъявляются к корпусам, которые разрабатываются для высококачественных систем автоматического распознавания речи и озвучивания текстов. Каждый корпус создаётся с определённой целью, которая учитывает определённую специфику научных исследований или разрабатываемых прикладных систем.

Создание данного корпуса длится уже около двух лет. Результат этой работы — пилотная версия корпуса эфирной речи.

Цель данной работы — описание структуры корпуса, средств формирования корпуса, первых конкретных результатов анализа и использования речевого материала, а также перспектив дальнейших исследований.

Опыт создания речевых корпусов в Украине

Речевой корпус состоит из структурированного множества речевых фрагментов, описания этих фрагментов, а также компьютерных средств для оперирования со всем множеством данных корпуса.

Речевой фрагмент как базовая единица корпуса — это оцифрованный фрагмент речевого сигнала, который сопровождается ассоциированной информацией определённого типа (типов). Такая информация называется аннотацией речевого фрагмента [1].

Создание акустических корпусов — достаточно сложная научная и технологическая задача, которая требует значительных ресурсов. В 90-е гг. XX в. во многих странах были созданы координационные центры для сбора, хранения и распространения общедоступных и стандартизированных корпусов, в том числе и речевых [2]. Создание акустических корпусов становится самостоятельным направлением речевых технологий.

В Украине первые корпуса речи были созданы в 70-е гг. прошлого столетия для тестирования и оценки показателей систем распознавания речи на одинаковом стандартном речевом материале. Корпус из 1 тыс. отдельных слов использовался для тестирования системы распознавания на основе ЭВМ БЭСМ-6, при этом была достигнута точность распознавания в 96% при словаре в 1 тыс. слов [3]. Также для тестирования кооперативной (многодикторной) системы распознавания была накоплена выборка из 1600 реализаций из словаря в 100 слов для 6 дикторов. Было показано, что метод кооперативного обучения позволяет достичь 92% точности распознавания речи диктора, не входящего в кооператив [4].

В 90-е гг. XX в. для тестирования распознавания ключевых слов была записана английская слитная речь 11 дикторов длительностью в 3500 слов и размечена экспертами на отдельные слова, а часть материала — для обучения на отдельные фонемы [5].

Толчком в развитии пофонемного распознавания украинской речи послужил однодикторный корпус, содержащий более 6 тыс. изолированных слов, в значительной мере покрывающих фонетическое разнообразие языка. Акустическая модель, созданная на основе этого корпуса, позволила превысить надёжность 95% на словаре 3 тыс. слов. Также, начиная с 2004 года, успешно демонстрировалась базовая технология фонетического стенографа, как одно из достижений Государственной научно-технической программы «Образный компьютер».

Создание алгоритма распознавания речи из сверхбольших словарей (до 2 млн слов) потребовало накопление корпуса речи в 14 тыс. слов и сочетаний слов. Была достигнута

точность распознавания в 99,9% для словаря в 1 тыс. слов, а также точность в 85% для словаря в 2 млн слов при среднем времени распознавания в 7 сек. [6].

Многодикторный корпус «UkReco» содержит более 30 тыс. реализаций фонетически сбалансированных слов и фраз, записанных от около 100 дикторов из разных регионов Украины. Этот корпус используется для распознавания изолированных слов, адаптации на голос диктора, а также для построения акустических моделей для словаря-переводчика [7].

Другой размеченный корпус речи, записанной через телевизионную сеть, состоит из выступлений около 330 депутатов Верховной Рады Украины. Речь депутатов отличается быстрым темпом, спонтанностью и эмоциональностью. Объём обучающей выборки — 54 часов речи, контрольной — 11 часов речи. Средняя точность распознавания для контрольной выборки составила 71% [8].

Для исследования методов послогового и морфемного распознавания речи был накоплен корпус из более 35 часов читаемой речи одного диктора [9].

Интересное направление использования корпусов речи — их использование для синтеза речи. Такие корпуса предъявляют особые требования к качеству записи и подробности описания речевого сигнала. Для озвучивания украиноязычных текстов был записан женский голос профессионального диктора в студийных условиях [10].

Опыт, накопленный в предыдущих разработках, стал неоценимым при создании концепции данного корпуса эфирной речи.

Структура и состав акустического корпуса

Акустический корпус украинской эфирной речи (Акустичний корпус українського ефірного мовлення — АКУЕМ) — общий по цели своего применения акустический корпус, который содержит читаемую, подготовленную и спонтанную речь (последнее составляет самую большую часть корпуса). Все материалы корпуса по типу речевого сигнала разделяются на теле- и радиовещание, также присутствуют небольшие вкрапления записи публичной речи и речи в естественной среде. Основные языки материалов корпуса — украинский и русский.

В АКУЕМ вошли материалы разной тематики и жанров, но основу корпуса составили звуковые записи рубрик: новости и интервью (политика, культура, образование, общество и т.д.), телепередачи и телетрансляции (судебные заседания, политические дебаты, публичные выступления и др.). В целом корпус должен отображать полную картину речи украинского теле- и радиоэфира, поэтому работы над его пополнением будут вестись и в дальнейшем. В настоящее время количественное распределение звуковых записей по жанрам неравномерно. Это связано с первоочерёдностью отбора речевого материала определённой тематики, необходимой для работ по созданию системы распознавания речи (см. рис. 1).

На данный момент корпус украинской эфирной речи характеризуется следующими количественными показателями: более 260 часов аннотированной речи, словарь корпуса содержит почти 45 000 слов украинского языка и почти 50 000 слов русского языка, более 1500 тыс. дикторов. Среди записей встречается речь дикторов разного возраста, пола, социального положения и профессий, что отражает состав дикторов телевизионного эфира.

Кроме общеупотребительных слов, был создан словарь суржика (более 1700 слов), словарь территориальных и социальных диалектов (более 800 слов).

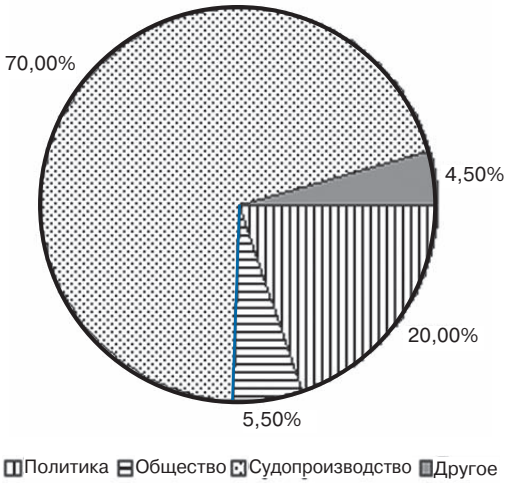

Рис. 1. Соотношение записей разных тематик (по продолжительности)

Разметка корпуса

Одна из основных черт, которые отличают акустический корпус от обычной коллекции звуковых записей или текстов, его разметка (аннотирование) — описание дополнительной информации о речевом сигнале.

Разметка АКУЕМ проводилась экспертами на основании предварительной автоматической разметки. Фактически, эксперт перепроверяет предварительную аннотацию, исправляя ошибки, делая необходимые дополнения, а также добавляя информацию о дикторах.

Разметка заключается в сегментировании речевого материала и детальном описании определённых лингвистических и экстралингвистических явлений в речевых фрагментах. Для внесения дополнительной информации в текст используются специфицированные теги, которые отделяются от текста знаком *. На данный момент используется 74 таких метаобозначения.

Все обозначения можно разделить на несколько групп:

- обозначения языка;
- обозначения нелитературных слов;
- обозначения способа произношения слов;
- обозначения фона;
- обозначения неинформативных слов и звуков, которые произносит диктор;
- обозначения диалогов и хоров;
- обозначения шума.

Обозначения языка касаются всех последующих слов до альтернативного обозначения, и ставятся перед первым словом соответствующего языка. В АКУЕМ в настоящее время встречаются записи девяти языков, хотя основной объём (более 97%) составляет украинский и русский языки.

Следующая группа обозначений предназначена для слов, отсутствующих в литературных словарях:

- суржик — смесь украинского и русского языков;
- социальные диалекты (жаргон, арго) — языки людей, связанных определённой общностью профессиональных или социальных интересов;
- территориальные диалекты — языки лиц, связанных между собой территориальной общностью;
- аббревиатуры и сокращения.

Обозначения способа произношения слов включают обозначения дефектов речи (заикание, картавость и др.), речевых сбоев (обрывы и оговорки), специфического произношения слов (например, послогового, с редуцированием или с растягиванием). Все эти обозначения касаются только одного слова и ставятся перед соответствующим словом.

Обозначения неинформативных слов и звуков, которые произносит диктор, включают обозначения заполненных пауз, звуков-паразитов и подобных явлений. К этой же группе относятся неинформативные звуки, например покашливание, шмыганье носом, смех, плач, громкий вдох или выдох диктора. Такие обозначения ставятся на месте соответствующего звука и обозначают соответствующие звуки в записи, которые произносит диктор. Эта группа обозначений самая большая.

Обозначения диалогов соответствуют местам в звуковых записях, где так или иначе сливается речь нескольких дикторов. Диалог — места, где во время разговора двух дикторов конец фразы первого диктора накладывается на начало фразы другого диктора. Хор — полное наложение речи нескольких дикторов.

Важная группа обозначений, которые описывают звуковые сегменты корпуса, — обозначения фона, на котором говорит диктор, и разнообразных шумов, которые присутствуют в сегментах. Такие обозначения касаются целого сегмента речи.

Примеры обозначений и частота их использования приведены в таблице 1.

Таблица 1

Некоторые обозначения, которые используются для описания сегментов АКУЕМ

Обозначение	Значение	Частота использования
у	украинский язык	3299
р	русский язык	3240
р-ак	русский язык с сильным иностранным акцентом	90
с	суржик	6300
ж	жаргон, арго	943
об	оговорка	3388
карт	картавость (неправильное произношение звуков «р» и «л»)	1191
см	смех	369
е	экание	15030
хор	хор	6066
пт	шелест бумаги (фон)	4695
мт	музыка (фон)	11718
опл	аплодисменты (фон)	1281
стук	стук	979
вул	шум улицы	2

Целевая аудитория АКУЕМ

Целевая аудитория проекта в первую очередь — разработчики систем автоматического распознавания украинской и русской речи. АКУЕМ предназначен для обучения и тестирования таких систем распознавания речи. Современ-

ным статистическим системам распознавания речи необходим большой объём акустических материалов для построения акустических и лингвистических моделей (далее АМ и ЛМ) речи, а также для тестирования надёжности распознавания речи.

На материалах корпуса проводятся многочисленные научные эксперименты в области распознавания речи, например, выявление и классификация экстралингвистических речевых явлений, исследование реальных акустических условий речи, исследование различных вариантов произношения дикторов, изучение специфики устной спонтанной речи на разных уровнях и много других.

АКУЕМ отображает современную языковую ситуацию в Украине, включает как литературный, так и разговорный стиль речи. Поэтому корпус может служить основой для широкого спектра исследований в области лингвистики, диалектологии, речевой акустики, психоакустики, фонетики, фонологии и других областях науки.

Программное обеспечение корпуса

Эффективное создание корпуса невозможно без развитого инструментария. К этому инструментарию относятся программные средства для стенографирования звуковых записей, дальнейшего их сегментирования и аннотирования (транскрибирования), автоматического исправления транскрипций, статистического анализа результатов сегментирования, а также подготовки материала к обучению распознавания.

Средства стенографирования звукозаписей

Стенографирование производится средствами протоколирования событий SRS-Femida [11].

Стенографист создаёт транскрипцию звукозаписи с уровнем детализации, которая включает признаки языка и говорящего. С помощью ножной педали осуществляется предварительное разделение на речевые сегменты, которые отвечают смене говорящего. Видеоряд, который сопровождает звукозаписи, облегчает определение диктора.

Кроме этого, указываются участки сигнала, где речь неразборчивая, перекрывается шумами или отсутствует.

Для обеспечения орфографической правильности набранного текста используются стандартные средства проверки орфографии, адаптированные к специфике стенограмм: учитываются обозначения языковых признаков и добавляются признаки отклонения от нормативов литературного языка для соответствующих слов.

Средства сегментирования

Сегментирование выполняется средствами программного обеспечения с открытым кодом *Transcriber 1.5.1* [12] (см. рис.2), адаптированного к кириллице. Подготовленный специалист с соответствующим уровнем лингвистического и компьютерного образования углубляет детализацию транскрипции, полученной в результате стенографирования звукозаписей. Проводится тщательное разбиение по паузам речевых сегментов, синхронизация их с соответствующим текстом. Кроме этого, в текст вставляются детальные признаки-теги, которые касаются как отдельных слов и звуков, так и речевого сегмента в целом.

Дальнейший анализ сегментирования состоит в выявлении и исправлении типичных ошибок и внесении некоторых регулярных изменений, обусловленных как непрерывным развитием концепции корпуса, так и появлением различных версий использования корпуса.

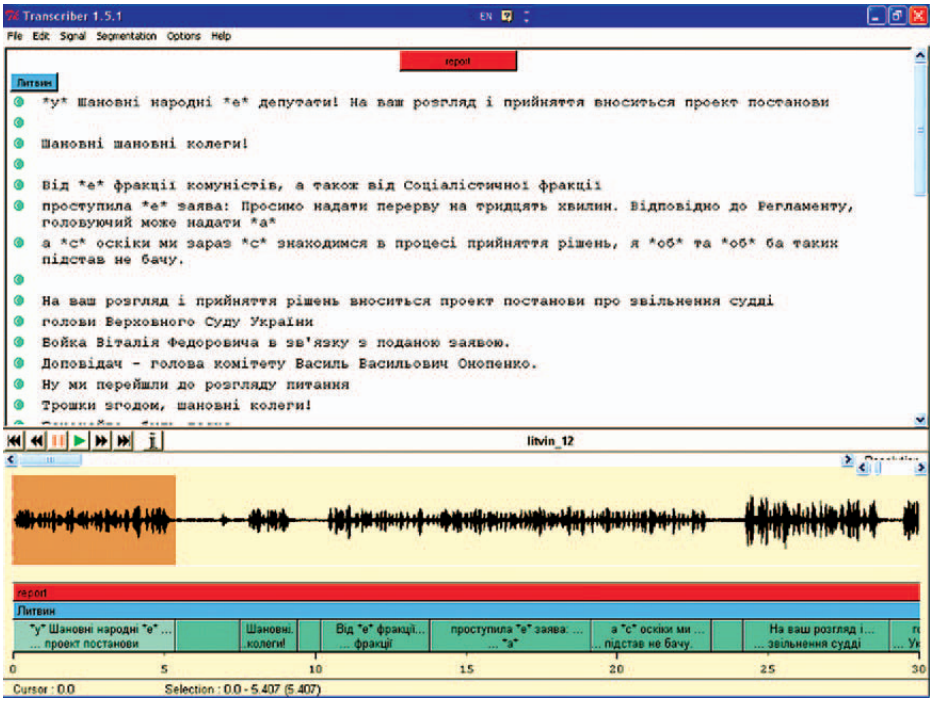

Рис. 2. Диалоговое окно эксперта в программе Transcriber

Средства статистического анализа

Для анализа накопленного материала производится подсчёт различных статистик, в частности формируются:

- частотные словари для разных языков, которые встречаются в корпусе;
- частотные словари суржика, социальных и региональных диалектов, аббревиатур, редуцированных слов и др.;
- статистика длин речевых сегментов для каждого звукового файла, а также общая статистика;
- статистика длин речевых сегментов для каждого диктора в отдельности.

Средства подготовки к обучению распознавания речи

Кроме указанного выше, производится формирование звуковых файлов, применимых для обучения и распознавания речи. При этом каждому звуковому фрагменту соответствует текстовая запись и имя диктора.

Словарь системы распознавания дополнен словами, которые отвечают неинформативным звукам (например, заполненные паузы) и, соответственно, во время формирования текста фразы эти звуки рассматриваются как отдельные слова. Были проведены эксперименты по обучению таким звукам-словам, и результаты показали высокую точность их определения (около 80%).

Обозначения, которые характеризуют целый сегмент, например, *стук*, *вул*, предлагается использовать для построения моделей гауссовских смесей (GMM) для того, чтобы система распознавания определяла такие сегменты и относила их к соответствующему классу.

Информацию о дикторе предлагается использовать для настраивания системы распознавания на кластеры дикторов. Это позволит повысить точность рас-

познавания за счёт предварительного определения кластера дикторов и использования индивидуальной акустической модели для данного кластера.

Предварительные эксперименты по распознаванию слитной речи

1. Речевой материал

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 1500 дикторов. Распределение неравномерное: большинство дикторов представлено короткими записями, однако, у 150 дикторов длительность записей составляет более 10 минут.

2. Текстовый материал

Текстовый материал, использованный для построения лингвистических моделей, состоит из текстов, загруженных из Интернета (400 Мбайт). Загруженный текст был модифицирован для того, чтобы убрать служебную информацию, записать числа в текстовом виде, а также отделить тексты на разных языках. В дополнение к этим текстам использовались также расшифровки звукового материала из обучающей выборки АКУЕМ.

3. Контрольная выборка

Для распознавания использовались записи длительностью 3,74 часа, в которых встретилось 29 500 слов. Всего в контрольной выборке присутствовала речь 34 дикторов. Темп произнесения — средний и быстрый.

4. Система распознавания речи

Для исследований использовался инструментарий HTK [13]. На его основе была создана многодикторная система распознавания речи.

В качестве АМ используются скрытые Марковские модели, обученные на обучающей выборке. 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.

В качестве лингвистической модели языка использовалась биграммная статистическая модель.

Словарь распознавания, используемый наряду с уже обученными ЛМ и АМ, насчитывал 42 598 словоформ. Произнесение каждой словоформы было 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»).

Результаты распознавания приведены в таблице 2. Заметим, что в контрольной выборке наряду с записями телепередач присутствуют записи реального судьи Ш.

Таблица 2

Результаты распознавания речи

Дикторы	Профессия	Надёжность распознавания (%)
Окис	актёр в роли судьи	73,47
Калинская	актриса в роли судьи	58,65
Ш.	судья	59,47
Антонюк	актриса в роли прокурора	63,90
Наум	актёр в роли прокурора	59,10
Бойко	актёр в роли прокурора	57,76
Бевз	актёр в роли адвоката	55,93
Жуковская	актриса в роли адвоката	66,38
Бабич	актриса в роли адвоката	51,64
Бузаджи	актёр в роли адвоката	60,28
Солодко	актёр в роли адвоката	46,95
Сологуб	актриса в роли судебного секретаря	81,26
В среднем		59,61

Выводы

Разработанная пилотная версия AKYEM позволяет строить акустические и дополнять лингвистические модели для исследования по автоматическому транскрибированию звуковых сигналов, для поиска ключевых слов, а также для распознавания дикторов.

Дальнейшие исследования предусматривают построение информационно-поисковой системы на основе веб-интерфейса, который позволит пользователям ориентироваться в речевом материале и находить в нём нужную информацию более эффективно. Также полезными могут оказаться средства для синхронизации текстовых и речевых материалов.

Несмотря на сложность и трудоёмкость, мы надеемся создать полноценный ресурс, который станет основой для многих речевых технологий и систем, которые могут использоваться во многих сферах экономики, образования, права и повседневной жизни. Материал корпуса состоит из разнообразных звуковых записей вместе с их расшифровкой и может стать частью Национального корпуса украинского языка.

Литература

1. *Кривнова О.Ф.* Речевые корпуса на новом технологическом витке // Речевые технологии. 2008. № 2. С. 13–24.

2. *Кривнова О.Ф., Захаров Л.М., Строкин Г.С.* Речевые корпусы (опыт разработки и использование) // Труды семинара Диалог'2001. Москва, 2001.

3. *Винцюк Т.К., Шинкаж А.Г.* Распознавание 1000 слов // Автоматическое распознавание слуховых образов. Тбилиси, Мецниереба, 1978.

4. *Винцюк Т.К., Куляс А.И., Людовик Е.К., Шинкаж А.Г.* Кооперативная система распознавания речи // Автоматическое распознавание слуховых образов. Ереван, 1980.

5. Вінцюк Т., Біднюк С., Куляс А., Пилипенко В., Дослідження з розпізнавання ключових слів у потоці зв'язного мовлення // Праці першої всеукраїнської конференції УкрОБРАЗ 92. Київ, 1992. С. 125–128.

6. *Pylypenko V*. Information Retrieval Based Algorithm for Extra Large Vocabulary Speech Recognition // Proc. of the 13th International Conference «Speech and Computer: SPECOM'2006». St. Petersburg, Russia, 2006. P. 67–69.

6. *Сажок М., Селюх Р., Юхименко Ю*. Адаптація акустичних моделей фонем до голосу диктора для пофонемного розпізнавання ізольованих слів української мови // Штучний інтелект. Донецьк, 2009. № 4. С. 230–233.

7. *Pylypenko V., Robeiko V*. Experimental System of Computerized Stenographer for Ukrainian Speech. // Proc. of the 13th International Conference «Speech and Computer: SPECOM'2009». St. Petersburg, Russia, 2009. P. 67–70.

8. *Васильєва Н., Сажок М*. Порівняння пофонемного та поскладового розпізнавання мовленнєвого сигналу для української мови // Праці десятої всеукраїнської міжнародної конференції УкрОБРАЗ, Київ, 2010. С 49–54.

9. *Lyudovyk T., Brozinski S., Noner M., Robeiko V., Sazhok M*. Speech Synthesis Applied to SMS reading // Proc. of the 13th International Conference «Speech and Computer: SPECOM'2009». St. Petersburg, Russia, 2009. P. 300–305.

10. *Радуцький О., Богданов Л*. Технічна фіксація судових процесів: системний підхід до розвитку комп'ютерних технологій та інформаційних ресурсів // Юридичний журнал. 2002. № 2. http://www.justinian.com.ua/article.php?id=431

11. *Barras C., Geoffrois E., Wu Z., Liberman*. Transcriber: a free Tool for Segmenting, Labeling and Transcribing Speech. In: Proc. First Int. Conf. on Language Resources and Evaluation (LREC 98), Granada, Spain, M., 1998. P. 1373–1376.

12. *Young S*. et al. The HTK Book (for HTK Version 3.4) // Cambridge University Engineering Department: Cambridge, UK, 2009. http://htk.eng.cam.ac.uk/

Сведения об авторах

Васильева Нина Борисовна —
научный сотрудник отдела распознавания и синтеза звуковых образов Международного научно-учебного центра информационных технологий и систем, г. Киев, Украина. n.vassilleva@gmail.com

Пилипенко Валерий Васильевич —
старший научный сотрудник отдела распознавания и синтеза звуковых образов Международного научно-учебного центра информационных технологий и систем, г. Киев, Украина. valeriy.pylypenko@gmail.com

Радуцкий Александр Михайлович —
кандидат технических наук, директор ООО «Специальные регистрирующие системы», г. Киев, Украина. alex@srs.kiev.ua

Робейко Валентина Васильевна —
научный сотрудник отдела распознавания и синтеза звуковых образов Международного научно-учебного центра информационных технологий и систем, г. Киев, Украина. valya.robeiko@gmail.com

Сажок Николай Николаевич —
кандидат технических наук, старший научный сотрудник отдела распознавания и синтеза звуковых образов Международного научно-учебного центра информационных технологий и систем, г. Киев, Украина. sazhok@gmail.com