ВЛИЯНИЕ РЕЧЕВЫХ ХАРАКТЕРИСТИК ДИКТОРОВ НА ТОЧНОСТЬ АВТОМАТИЧЕСКОГО РАСПОЗНАВАНИЯ РЕЧИ

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

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

Исследуется влияние речевых сбоев, темпа речи и других характеристик на точность распознавания украинской подготовленной и спонтанной речи. Обсуждается возможность предсказания точности распознавания речи дикторов исходя из их опыта публичных выступлений. Предлагается область применения системы распознавания речи.

Ключевые слова: автоматическое распознавание речи, украинский язык, спонтанная речь, речевые характеристики

EFFECT OF SPEECH AND SPEAKER CHARACTERISTICS ON ACCURACY OF AUTOMATIC SPEECH RECOGNITION

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

International Research/Training Center for Information Technologies and Systems, Kyiv, Ukraine

This paper reports an investigation of effects caused by speech style and speaker characteristics on speech recognition accuracy. Results of Ukrainian read and spontaneous speech recognition are analized.
The speech material consists of broadcast news (15 %), talk shows (29 %), and real court reports (31 %). The test corpus counts 17 000 words and includes speech of 9 male and 9 female speakers. The speakers are TV news presenters, politicians, journalists, lawyers, and "ordinary" members of trials.
Two experiments have been conducted. The first one consists in automatic speech recognition, while the second one is based on annotations of speech made by experts.
Different speech characteristics are investigated, namely speaking rate, speech disfluences (breathing, hesitation fillers, restarts, fragmented words, reduction and prolongation of words), slang and colloquial words. The most error-prone speech characteristics are restarts and fragmented words, slang and colloquial words. Breathing and hesitation fillers are less error-prone because they are successfully modeled in the speech recognition system. Fast read speech is recognized more accurately than more slow spontaneous speech of the same speaker.
The more accurate recognition (word error rate 10–15 %) is achieved for speakers who have experience in oral presentations. The obtained results allow to roughly predict the accuracy of speech recognition based on such speaker experience. Ukrainian speech recognition accuracy achieved so far allows to use the speech recognition system for automatic transcription (subtitling) of broadcast news.

Key words: automatic speech recognition, Ukrainian language, spontaneous speech, speech characteristics.

1. Введение

Дикторонезависимое распознавание слитной речи с большим словарем (speaker independent large vocabulary continuous speech recognition) позволяет решать широкий спектр задач, таких как транскрибирование (субтитрирование, стенографирование) аудиозаписей телепередач [1], заседаний парламента [2], судебных заседаний [3] и т. п. Сложность этих задач обусловлена тем, что

распознаваемая речь может быть как подготовленной (заранее обдуманное выступление), так и неподготовленной (чтение «с листа», спонтанная речь, например, подсудимых, свидетелей и т. д.). В некоторых случаях распознаваемая речь может быть эмоционально окрашенной (выступления некоторых политиков, а также речь адвокатов и других участников судебных заседаний). Иногда фрагменты подготовленной речи чередуются со спонтанными диалогами и полилогами.

Для решения подобных задач системы распознавания речи должны быть устойчивы не только к техническим факторам (шумовые помехи, свойства каналов передачи акустических сигналов), но и к речевым сбоям, вызываемым ситуацией и индивидуальными речевыми особенностями дикторов [4–7]. К последним относятся оговорки, запинки, паузы хезитации, изменение темпа речи, а также нелитературная лексика (в первую очередь имеется в виду суржик и сленг). Кроме этого, в речи многих дикторов присутствуют такие паралингвистические явления как вдох/выдох, кашель, смех, причмокивание и т. п.

Подготовленная речь (например, новости, читаемые дикторами телевидения) в настоящее время может распознаваться с пословной точностью 85–90 % [8]. Для этого требуется предварительное обучение системы распознавания на материале больших речевых корпусов (от нескольких сотен до нескольких тысяч часов записанной речи). Распознавание спонтанной речи — значительно более сложная задача, требующая не только соответствующего материала для обучения, но и учета особенностей такой речи при разработке акустико-фонетических и лингвистических моделей систем распознавания.

Анализ проблем, связанных с особенностями спонтанной речи [9], показал, что одной из основных причин ошибок распознавания являются речевые сбои, к которым относятся оговорки, паузы хезитации, а также слишком акцентированное и/или растянутое произнесение отдельных слов. При этом наличие в речи оговорки привело к ошибке распознававния в 100 % случаев, паузы хезитации — в 92 %, растянутого произнесения — в 81 %.

Разработчики из LIMSI (Франция) провели эксперимент со своей системой распознавания речи [1], целью которого было выявить степень влияния речевых сбоев на точность распознавания. Материалом для эксперимента послужили записанные спонтанные телеинтервью на французском языке. Точность распознавания составила 85.5 %. Речевые сбои явились причиной ошибки в 12.5 % случаев. Основной причиной ошибок (25.1 %) явилось редуцированное произнесение слов, свойственное спонтанной речи.

Темп речи, несомненно влияющий на точность распознавания, а также косвенно связанный с речевыми сбоями, исследовался в [9, 10]. В целом, темп спонтанной речи варьирует больше, чем темп подготовленной речи.

Связь речевых сбоев с возрастом и полом дикторов исследовалось на материале английского [4] и венгерского [11] и языков. Подтвердилось предположение о том, что в детском и пожилом возрасте плавность речи нарушается чаще, чем в среднем возрасте. Что касается различия полов, то вопрос о том,

кто более склонен к речевым сбоям — мужчины или женщины — остается открытым.

Процентное содержание речевых сбоев в речи может сильно колебаться: от 1.33 % ([12]) до 5.97 % ([4]).

Интересные результаты были получены при исследовании речи, обращенной к компьютеру [12]. Эксперимент показал, что в речи людей меньше сбоев наблюдается при общении с компьютером (0.78 % — 1.87 %), чем при общении с другими людьми (5.50 % — 8.83 %). Возможно, это связано с более медленным темпом, а также большей мотивацией продумать (спланировать) реплику от начала до конца. По крайней мере, результаты данного исследования говорят о том, распознавание речи в человеко-компьютерных системах должно быть не сложнее, чем распознавание спонтанной речи, адресованной людям.

2. Цель исследования

Целью работы является исследование влияния ситуативных и индивидуальных факторов на точность распознавания речи. Под ситуативными факторами в данной работе понимаются условия протекания речевой коммуникации (теледебаты, интервью, чтение теленовостей, судебный процесс) и роли коммуникантов. К индивидуальным факторам отнесены степень подготовленности к публичным выступлениям, склонность к речевым сбоям, степень употребления нелитературной лексики.

Полученные результаты могут быть использованы для предсказания точности распознавания речи при появлении новых дикторов и/или новых приложений системы.

Полученная информация позволит определить сферы применения системы распознавания речи исходя из результатов, достигнутых на текущий момент.

Предполагается определить степень влияния различных индивидуальных характеристик дикторов на точность распознавания их речи. В частности, внимание уделяется речевым сбоям, темпу произношения, паралингвистическим явлениям и употреблению нелитературной лексики (суржика).

Конечной целью работы является повышение точности распознавания как подготовленной, так и спонтанной речи.

3. Речевой материал

Исследованный речевой материал, содержащий 16 699 реализаций словоформ, был взят из Акустического корпуса украинской эфирной речи (АКУЕМ) [13] (86 %), непосредственно из телеэфира (11 %) и записан в студии (3 %). Тип сигнала — моно, частота дискретизации 22,05 кГц с 16-битным разрешением, формат Windows PCM wav.

Аудиозаписи корпуса АКУЕМ, разработанного фирмой «Специальные регистрирующие системы» (Киев, Украина), имеют аннотацию, включающую орфографический текст и сведения о речевых сбоях (оговорки, паузы хезитации), паралингвистеских явлениях (вдох/выдох, кашель, смех, плач), а также о значительно редуцированном или растянутом произнесении. Кроме этого, в корпусе АКУЕМ отмечены такие явления как суржик, жаргон, диалектизмы.

По содержанию исследованный материал представляет собой новостные телепередачи (15%), ток-шоу (29%), телепередачи судебной тематики (25%), а также речь, прозвучавшую в реальных судебных заседаниях (31%). Таким образом, представлена подготовленная речь (чтение новостей), слабо подготовленная речь (выступления политиков, обвинительные речи прокуроров), неподготовленное чтение (зачитывание судебных постановлений) и спонтанная речь (в основном, опросы свидетелей и судебные прения).

Исследовалась речь дикторов телевидения, политиков, журналистов, юристов и «рядовых» участников судебных процессов. Женщин и мужчин было поровну. В Таблице 1 приведены сведения о дикторах. Заглавная буква кода соответствует роду занятий диктора или его роли в речевом общении (Ю — юрист, П — политик, Д — диктор телевидения, Ж — журналист, С — свидетель (возможно также подсудимый или потерпевший) в телепередаче судебной тематики. Следует подчеркнуть, что в роли судей, прокуроров и адвокатов выступают реальные юристы. Строчная буква кода обозначает пол диктора.

Особый интерес представляет речь юриста Ю_ж_1, поскольку исследовались записи реальных судебных заседаний с ее участием. В одних случаях судья опрашивала подсудимых и свидетелей (спонтанная речь), в других — оглашала протоколы заседаний (чтение).

Таблица 1. Распределение дикторов в соответствии с их родом деятельности и полом

Коды дикторов	**Д_ж**	**Д_м**	**П_ж**	**П_м**	**Ж_ж**	**Ж_м**	**Ю_ж**	**Ю_м**	**С_ж**	**С_м**	**Всего**
Количество дикторов	2	1	1	5	2	1	2	1	2	1	18

4. Система распознавания украинской речи

Система распознавания украинской речи [2, 3] разработана на основе инструментария HTK [14].

В качестве акустических моделей (АМ) используются скрытые Марковские модели 70 фонем, обученные на материале корпуса АКУЕМ (82 часа аннотированной речи, 1150 дикторов). Для обучения был отобран речевой материал, не содержащий таких помех как одновременное звучание нескольких голосов, звук аплодисментов и другие не зависящие от диктора шумы. Помимо акустических моделей 53 фонем украинского языка используются особые модели для

паузы, звучащих пауз хезитации («э-э-э», «а-а-а», «м-м»), вдохов/выдохов, чмоканья, кашля, смеха и плача.

Лингвистическая модель (ЛМ) представляет собой биграммную модель языка, заданную вероятностями пар словоформ. ЛМ обучена на текстах корпуса АКУЕМ, текстах из Интернета и нескольких искусственно созданных текстах, в которых представлены все комбинации «день+месяц» и часть комбинаций «день+месяц+год» (например, в переводе с украинского: «первое января», «первого января», «второго февраля две тысячи двенадцатого года»).

Словарь распознавания содержит 115 900 словоформ. Часто употребляемые в речи словоформы, в том числе числительные, имеют от 1 до 10 фонемных транскрипций, остальные — от 1 до 3. В среднем на одну словоформу в словаре распознавания приходится 1.55 фонемных транскрипций, отражающих как типичное, так и спонтанное произнесение. Основная часть транскрипций порождена автоматически. Так, 3924 транскрипции числительных были созданы автоматически, а 34 добавлены вручную (ср. русское «тысячи» [т Ы ш’ и]).

5. Эксперименты

Было проведено два эксперимента, один из которых состоял в распознавании речи, а второй основывался на корпусном исследовании.

5.1. Распознавание речи

С помощью системы распознавания украинской речи были получены данные о точности распознавания речи 18 дикторов. Пословная точность распознавания вычислялась с помощью инструментария НТК путем сравнения «правильного» (эталонного) текста (reference) с ответом распознавания в виде последовательности словоформ. Учитывались замены, вставки и пропуски словоформ. Не учитывались вдохи/выдохи и хезитации.

5.2. Корпусное исследование

На материале корпуса АКУЕМ были исследованы речевые характеристики, зависящие от диктора и ситуации речевого общения — речевые сбои, темп, редуцированное и растянутое произнесение отдельных слов, а также присутствие в речи нелитературной лексики. Речевой материал, не входящий в корпус, был предварительно аннотирован в соответствии со стандартами корпуса.

Темп речи измерялся количеством произнесенных в секунду слогов. Для каждого диктора экспертом были выбраны характерные участки речи без внутренних пауз. Подсчет слогов производился по гласным звукам (являющимся слогообразующими) соответствующих фонемных транскрипций, полученных

автоматически, без учета особенностей произношения дикторов. Полученные значения темпа речи рассматриваются как приблизительные, поскольку и гласные в речевом потоке могут выпадать, и темп в целом может варьировать в зависимости от таких факторов как усталость, переход от чтения к спонтанной речи и т.д.

Речевые сбои были объединены в две категории: хезитации («а-а-а», «м-м», ...) и оговорки (фальстарты и произнесение не того слова, которое планировалось).

Из нелитературной лексики в исследованном материале присутствовал только суржик.

Количественной оценкой речевых сбоев, суржика, а также вдохов/выдохов, редуцированного и растянутого произнесения отдельных слов являлось их процентное отношение к общему количеству слов исследуемого речевого материала диктора.

Полученные результаты были сопоставлены с результатами распознавания и представлены в Таблице 2. Речь диктора Ю_ж_1 представлена как Ю_ж_1(ч) и Ю_ж_1(с) — соответственно как чтение и спонтанная речь.

Таблица 2. Результаты распознавания речи и речевые характеристики дикторов

Дикторы	Количество слов	Точность распознавания	Темп речи	Шумное дыхание	Хезитация	Оговорки	Редукция	Растягивание	Суржик
Д_ж_1	968	89.77	7.7	0	0	0	0	0	0
П_м_1	736	88.32	5.7	0,8	1,5	0	0	0	0
Д_м_1	875	87.09	6.5	2,4	0,6	0,1	0	0	0
П_ж_1	276	84.06	7.3	0,7	0	0,7	0	0	0
Д_ж_2	459	83.66	6.0	0	0	1,5	0	0	0
П_м_3	491	81.06	6.2	0	0,6	0,6	0,2	1,6	1,2
С_ж_1	508	78.94	6.0	1,0	2,6	1,6	0,6	2,4	0
П_м_2	1258	77.90	5.6	0,6	1,3	0,5	0,7	3,1	0
Ю_ж_1(ч)	1553	76.11	7.4	5,6	0,1	0,6	1,5	0,3	0,4
Ю_ж_2	1976	74.80	4.9	1,0	3,0	1,5	0,8	1,0	0,5
Ж_ж_2	169	72.78	6.8	1,2	3,0	1,2	0	0,6	0
Ж_ж_1	255	71.76	6.6	3,9	0	0	1,6	0,4	0
Ю_м_1	1102	69.96	6.2	4,9	0,6	0,6	0,3	1,9	0,5
Ж_м_1	548	69.71	6.3	2,7	3,1	0	3,1	1,3	0
Ю_ж_1(с)	3725	64.35	5.3	5,3	2,8	0,6	0,1	0	2,5
П_м_4	315	62.86	7.3	1,0	0,3	0,3	3,2	1,3	0,3
П_м_5	958	57.83	6.8	0	2,0	0,3	2,6	1,0	0,3
С_м_1	175	47.43	4.7	1,1	5,7	1,7	0	12,6	4,6
С_ж_2	352	46.02	6.7	0	2,0	1,7	1,7	1,4	0,3

6. Обсуждение результатов

Эксперименты показали, что наименьшее влияние на точность распознавания оказывают шумные вдохи/выдохи и хезитации. Это объясняется тем, что данные явления успешно моделируются и распознаются системой [3]. Влияние темпа речи на точность распознавания неоднозначно. Быстрый темп речи дикторов телевидения, обладающих навыками четкой артикуляции, не ухудшает результаты, в то время как более эмоциональная быстрая речь политиков распознается хуже. На материале диктора Ю_ж_1 видно, что быстрая подготовленная речь (чтение) распознается на 12% лучше, чем медленная спонтанная.

В целом исследуемые речевые характеристики можно разбить на три группы: темп, характеристики, влияющие на точность распознавания (оговорки, редукция, растягивание, суржик) и не влияющие (вдохи/выдохи и хезитации). Эти группы характеристик предсталены на диаграмме (Рис. 1). Горизонтальная ось соответствует списку дикторов, упорядоченному по убыванию точности распознавания речи. По вертикали представлены значения точности (%), а также масштабированные с коэффициентом 10 значения темпа речи и с коэффициентом 5 суммированные значения влияющих и не влияющих на точность распознавания речевых характеристтик.

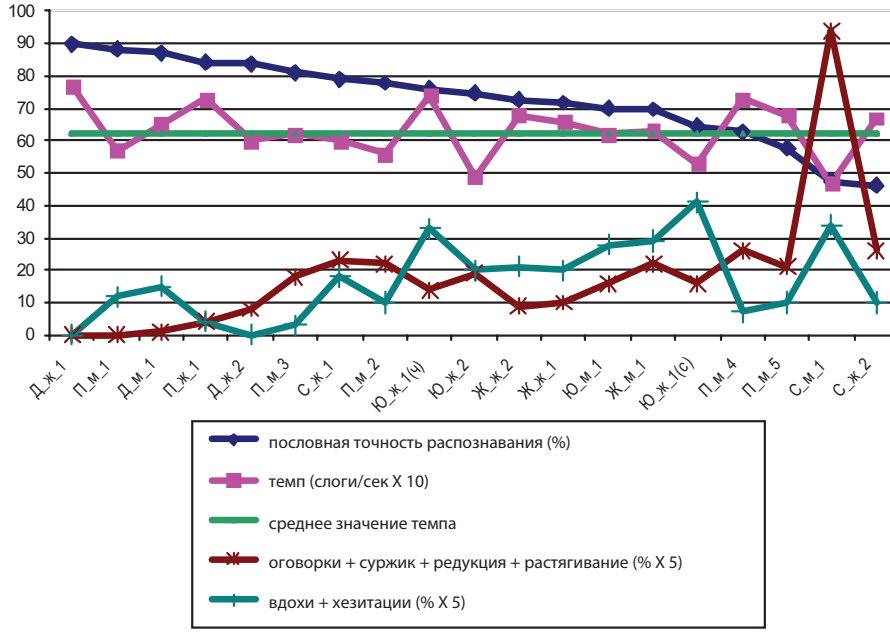

Рис. 1. Точность распознавания речи различных дикторов и их речевые характеристики

Особенно явно различия между группами характеристик заметны на результатах дикторов, чья речь распознана с точностью более 85% (первые три диктора) и с точностью менее 50% (два последних диктора). Речь «отличников», среди которых два диктора телевидения (женщина и мужчина), практически лишена суржика и оговорок. Речь «отстающего» диктора С_м_1 отличается медленным темпом и растянутым произнесением каждого восьмого слова.

Если рассматривать влияние речевых характеристик на точность распознавания речи, то основными причинами ошибок являются: 1) оговорки; 2) использование диктором нелитературной лексики; 3) редуцированное произнесение слов; 4) растянутое произнесение слов. Первые две причины приводят к тому, что «не срабатывает» лингвистическая модель системы распознавания, основанная на статистических закономерностях употребления слов. Третья и четвертая причины вызывают проблемы, связанные с акустической моделью системы распознавания речи.

Результаты распознавания украинской речи подтверждают вывод, сделанный нами ранее на материале распознавания русской речи: хорошо распознается речь дикторов, имеющих большой опыт публичных выступлений. Закономерно, что лучшие результаты по точности распознавания у дикторов телевидения, а также политиков, имеющих опыт работы пресс-секретарем, главой администрации, спикером. Причем, если в исследуемом материале речь профессиональных дикторов — это чтение текста, то речь политиков — речь не «по бумажке» (в основном ответы на вопросы).

Полученные результаты позволяют приблизительно предсказать точность распознавания речи диктора исходя из его профессии: профессиональные дикторы радио и телевидения — 80–90%, политики-«спикеры» — 75–90%, журналисты, телеведущие и юристы — 60–80%, остальные — 45–80%.

Важна также роль диктора в речевой ситуации. Предварительный анализ речи расширенного состава участников судебных заседаний свидетельствует о том, что в целом по точности распознавания речи прокуроры уступают судьям, но превосходят адвокатов. Это может быть объяснено, в частности, тем, что речь судей наиболее нейтральна, а речь адвокатов наиболее эмоциональна, и, следовательно, менее контролируема (в большей степени подвержена речевым сбоям), однако одновременно и более акцентирована на отдельных участках (растягивание слов, послоговое произнесение).

Немаловажным является мотивированность речи, желание диктора, чтобы его слушали и понимали. В этой связи можно упомянуть получающие все большее распространение речевые технологии голосового поиска, перевода, ввода текстов для электронной почты и т. п. Если есть заинтересованность в адекватной реакции персонального голосового помощника, с которым происходит устное общение, то необходимо произносить голосовые запросы и команды четко (но без гиперартикуляции) и в стабильном темпе. Можно предположить, что голосовое общение с техническими устройствами будет способствовать совершенствованию речевых навыков (своего рода «персональный логопед»).

Исследования, проведенные в данной работе, будут в дальнейшем продолжены в направлении автоматизации выявления значений речевых характеристик и совершенствования подходов их моделирования в системе распознавания речи.

7. Выводы

Точность распознавания речи зависит от того, насколько успешно моделируются в системе распознавания речи характеристики, обусловленные ситуативными и индивидуальными факторами.

Основными речевыми характеристиками, влияющими на точность распознавания речи, являются: 1) оговорки; 2) использование диктором нелитературной лексики; 3) редуцированное произнесение слов; 4) растянутое произнесение слов.

Фактором, положительно сказывающимся на точности распознавания речи, являются навыки устных выступлений. Полученные результаты позволяют приблизительно предсказать точность распознавания речи диктора исходя из его профессии.

Точность распознавания украинской речи, достигнутая к настоящему времени, позволяет использовать систему распознавания речи для автоматического субтитрирования новостных телепередач.

Литература

1. *Adda-Decker M., Habert B., Barras C., Adda G., Boula de Mareuil P., Paroubek P.* A disfluency study for cleaning spontaneous speech automatic transcripts and improving speech language models. Proceedings of DiSS'03, Disfluency in Spontaneous Speech Workshop, Göteborg, Sweden, 2003, pp. 67–70.
2. *Пилипенко В. В., Робейко В. В.* Автоматизированный стенограф украинской речи // Искусственный интеллект. Донецк: 2008. № 4. С. 768–775. Pilipenko V. V., Robeiko V. V. (2008), Automated stenographer of Ukrainian speech [Avtomaticheskii stenograf ukrainskoi rechi], Iskusstvennyi intellekt [Artificial Intelligence], no. 4, pp. 768–775.
3. *Людовик Т. В. Пилипенко В. В., Робейко В. В.* Автоматическое распознавание спонтанной украинской речи (на материале акустического корпуса украинской эфирной речи) // Компьютерная лингвистика и интеллектуальные технологиии. По материалам ежегодной Международной конференции «Диалог 2011». Вып. 10. М.: РГГУ, 2011. С. 478–488. Lyudovyk T. V., Pilipenko V. V., Robeiko V. V. Automated stenographer of Ukrainian speech (on material of acoustic corpus of Ukrainian speech) [Avtomaticheskoe raspoznavanie spontannoi ukrainskoi rechi (na materiale akusticheskogo korpusa ukrainskoi efirnoi rechi], Komp'iuternaia Lingvistika i Intellektual'nye Tekhnologii: Trudy Mezhdunarodnoi Konferentsii "Dialog 2011" [Computational

Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialog 2011"]. Bekasovo, 2011, pp. 478–488.

4. *Bortfeld H., Leon S. D., Bloom J. E., Schober M. F., Brennan S. E.* (2001), Disfluency Rates in Conversation: Effects of Age, Relationship, Topic, Role, and Gender, Language and Speech, Vol. 44 (2), pp. 123–147.
5. *Meyer B. T., Brand T., Kollmeier B.* (2011), Effect of speech-intrinsic variations on human and automatic recognition of spoken phonemes, Journal of the Acoustical Society of America, Vol. 129, Issue 1, pp. 388–403.
6. *Choularton S.* (2007), "Early Stage Detection of Speech Recognition Errors", available at: http://www.xesoftware.com.au/ThesisAsPassed.pdf
7. *Greenberg, S., Chang, S.* Linguistic dissection of switchboard-corpus automatic speech recognition systems. Proceedings of the ISCA Workshop on Automatic Speech Recognition: Challenges for the New Millennium. Paris, 2000, pp. 195–202.
8. *Le* Laboratoire d'Informatique pour la Mécanique et les Sciences de l'Ingénieur (LIMSI) (2010), Rapport d'activité 2010, available at: http://www.limsi.fr/Rapports/LIMSI_Rapport2010.pdf.
9. *Butzberger J. W., Murveit H., Shriberg E. E., Price P. J.* Spontaneous speech effects in large vocabulary speech recognition applications. Proceedings of the DARPA Speech and Natural Language Workshop. San Mateo, CA, 1992, pp. 339–343.
10. *Trouvain J., Koreman J., Erriquez A., Braun B.* Articulation rate measures and their relation to phone classification in spontaneous and read German speech. Proceedings of ITRW on Adaptation Methods for Speech Recognition Adaptation-2001. Sophia Antipolis, France, 2001, pp. 155–158.
11. *Menyhart K.* Age-dependent types and frequency of disfluencies. Proceedings of DiSS'03, Disfluency in Spontaneous Speech Workshop. Goteborg University, Sweden, 2003, pp. 45–48.
12. *Oviatt S.* (1995). Predicting spoken disfluencies during human-computer interaction. Computer Speech and Language, no. 9, pp. 19–35.
13. *Pylypenko V., Robeiko V., Sazhok M., Vasylieva N., Radoutsky O.* Ukrainian Broadcast Speech Corpus Development. Proceedings of SPECOM 2011, 14-th International Conference "Speech and Computer". Kazan, Russia, 2011, pp. 435–440.
14. *Young S.* et al. (2009), "The HTK Book (for HTK Version 3.4)", available at: http://htk.eng.cam.ac.uk/.