УДК 621.3

О влиянии помех и параметров настройки на качество функционирования системы автоматического распознавания речи

С. А. Ёлкин, **А.Г. Ляшенко**, **В.П. Овсяник**, д-р техн. наук, **А. Н. Продеус**, д-р техн. наук
Национальний технический университет Украины "Киивський политехничний институт",
пр. Победы, 37, Киев-56, 03056, Украина.

Экспериментально исследовано влияние помех и некоторых характеристик системы автоматического распознавания речи на качество распознавания. Получены рекомендации по оптимизации параметров системы автоматического распознавания речи для нескольких сценариев ее использования.

Ключевые слова: *распознавание речи, помехи, шумоподавление, параметризация, транскрибирование, речевой корпус.*

Введение

Развитию направления речевых технологий, за более чем 50 лет исследований, содействовали работы многих специалистов [3]. Одним из таких направлений является построение систем автоматического распознавания речи (САРР). Актуальность данного направления обусловлена широким распространением цифровых технологий обработки сигналов, на базе которых осуществляется автоматизация процессов анализа и управления большими и сложными системами различного назначения. Необходимость участия человека в таком анализе и управлении требует создания эффективных человеко-машинных интерфейсов. Кроме того, человеко-машинный интерфейс необходим людям с ограниченными возможностями.

Благодаря существованию программного инструментария Hidden Markov Model Toolkit (HTK) [2], создание компьютерных моделей САРР, а также их экспериментальные исследования, стали доступны широкому кругу пользователей. Примерами исследований такого рода, проведенных студентами технического вуза, могут служить работы [8, 9], где испытания возможностей инструментария НТК осуществлялись путем экспериментального поиска минимального размера вектора кепстральных признаков. Кроме того, в указанных работах затронуты вопросы выбора словаря фонем украинской речи, состава вектора кепстральных признаков, степени влияния объема обучающей выборки и уровня фонового шума на качество распознавания речи.

Цель данной работы состоит в продолжении линии, намеченной в работах [3, 4], а именно - в ней обобщены результаты, полученные независимо двумя студентами при иных испытаниях инструментария НТК. Эти результаты, а также способы их достижения, должны оказаться интересными и полезными для широкого круга исследователей и разработчиков САРР.

1. Постановка экспериментальных исследований

В рамках данной работы, двумя студентами, независимо друг от друга, проведены серии взаимодополняющих экспериментальных исследований компьютерных моделей САРР. В дальнейшем, рассматривая результаты первой серии экспериментов, следует понимать, что они получены одним из экспериментаторов, тогда как результаты другой серии экспериментов принадлежат второму экспериментатору.

В процессе исследований, помимо программного обеспечения в виде инструментария НТК [2], использовались звуковые редакторы Sony Sound Forge 10.0 [5] и Audacity 1.3.12-beta [1], с помощью которых производилась запись речевых сигналов, оценивалось отношение сигнал-шум, реализовывались процедуры подавления шумовой и импульсной помех.

Компьютерная модель САРР настраивалась на решение задачи распознавания дискретной речи (распознавание команд). При этом в режиме обучения САРР предъявлялись образцы одиночных слов, окруженных паузами длительностью не менее 0,3 с (согласно требованию инструментария НТК [2]). В режиме распознавания САРР предъявлялась единственная запись в виде всех слов речевого корпуса, разделенных паузами. Речевые корпусы (наборы звуковых файлов и соответствующих текстовых файлов-транскрипций) формировались самостоятельно каждым из экспериментаторов.

Одним из главных показателей качества работы САРР является точность (надежность) распознавания, измеряемая путем сравнения орфографических транскрипций контрольной вы-

борки (имеющихся в речевом корпусе), с результатом распознавания в виде последовательности словоформ. В данных экспериментальных исследованиях ошибками считались вставки, пропуски и замены слов.

2. Первая серия экспериментов

Исследовалась зависимость качества автоматического распознавания отдельных слов (речевых команд) от следующих характеристик речевого сигнала:

- фоновый шум и шумы диктора (шум дыхания, «щелчки» губ);
- количество образцов слов в обучающей выборке;
- состав словаря фонетических транскрипций слов.

Для проверки качества распознавания речи использовался самостоятельно сформированный речевой корпус украинской речи, содержащий записи произнесения 16-ти слов: резонанс; кожух; лінія; математика; кіт; піт; біт; плуг; луг; лук; пульт; функція; функцією; кристалічний; громадянин; громадяни. При записи сигналов использовались микрофон Logitech Analog Desktop Microphone (полоса частот 100…16000 Гц; чувствительность -67 дБ/мкбар) и звуковой редактор Sony Sound Forge 10 [5]. Параметры записанных в формате wav сигналов: 16 кГц; 16 бит; моно.

Влияние помех. Результаты распознавания для разного количества (1, 3, 5 и 10) образцов слов в обучающей выборке до и после применения процедур подавления помех приведены в табл. 1 и табл. 2, соответственно. В обоих случаях применялась MFCC_0_D_A-параметризация. Все операции по борьбе с помехами производились с использованием звукового редактора Sony Sound Forge 10.

Процедуры подавления помех применялись как для обучающей, так и для тестовой выборок. При этом подавление шумов диктора осуществлялось путем «ручного вырезания» шумов дыхания и «щелчков» губ, а подавление фонового шума производилось с помощью специального алгоритма, реализованного в программе Sony Sound Forge. Отношение сигнал-шум в исходных записях речевых сигналов составляло 19 дБ, применение процедур подавления помех позволяло улучшить этот показатель до значения в 35 дБ. Поскольку алгоритм шумоподавления является коммерческой тайной фирмы Sony, ограничимся указанием параметров этого алгоритма (рис. 1).

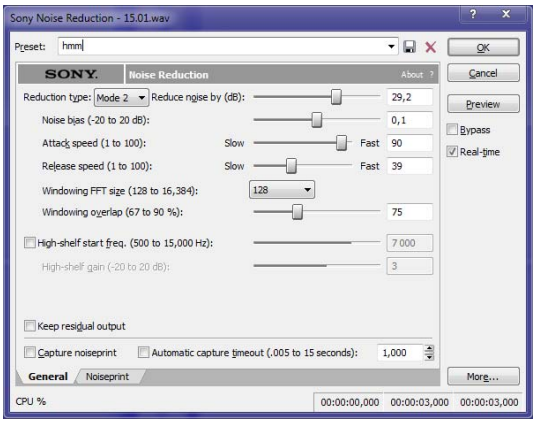

рис. 1.

Сопоставление результатов табл. 1 и 2 свидетельствует, что применение процедур подавления помех позволяет значительно улучшить качество распознавания. (Примечание: при суммировании содержимого строк приведенных таблиц, 100% должно получаться только для значений «правильно» и «замена»).

Таблица 1.

Кол-во вар. произн.	Правильно, %	Замена, %	Вставка, %
1	25	75	25
3	56	44	12
5	63	37	12
10	50	50	12

Таблица 2.

Кол-во вар. произн.	Правильно, %	Замена, %	Вставка, %
1	25	75	0
3	81	19	6
5	75	25	0
10	81	19	0

Зависимость качества распознавания от типа параметризации. Результаты исследований зависимости качества распознавания от выбора вида параметров САРР представлены в табл. 3. Отношение сигнал-шум в записях речевых сигналов составляло 19 дБ, а обучающая выборка содержала 10 образцов произнесения слов.

Таблица 3.

Тип параметризации	Правильно, %	Замена, %	Вставка, %
MFCC_0	94%	6%	6
MFCC_0_D	75%	25%	12
MFCC_0_D_A	69%	31%	19
MFCC_0_D_A_T	63%	37%	19

Приведенные в табл. 3 результаты свидетельствуют о том, что при распознавании зашумленной речи увеличение мерности признакового пространства является нецелесообразным. Оказывается, даже при сравнительно высоком отношении сигнал-шум 19 дБ наиболее простая MFCC_0-параметризация приводит к наилучшим результатам. Полученный результат нетривиален и представляет значительный интерес для разработчиков реальных САРР.

Заметим, что некоторые сведения о параметризации типа MFCC_0_D_A_T, не документированной в руководстве [2], можно найти в [4], где перечислены вообще все поддерживаемые HTK типы MFCC-параметризации.

Варьирование транскрипций. Была предпринята попытка «вручную» модифицировать автоматически сгенерированные транскрипции слов с целью получения более высокого качества распознавания. Для автоматического транскрибирования использовалась программа trans_rus_hgf.exe [7]. Ручная модификация транскрипций состояла в подборе более правдоподобных транскрипций, поскольку программа trans_rus_hgf.exe для некоторых слов генерировала явно не наилучшие транскрипции. Результаты этого эксперимента приведены в табл. 4. При этом в качестве обучающей и тестовой выборок использовались записи, подвергнутые процедурам подавления помех (табл. 2).

Таблица 4.

Кол-во вар. произн.	Правильно, %	Замена, %	Вставка, кол-во слов
1	12	88	6
3	75	25	0
5	81	19	0
10	87	13	0

Сопоставление табл. 2 и 4 свидетельствует, что редактирование результатов автоматического транскрибирования может приводить к заметному улучшению качества автоматического распознавания речи. В данном эксперименте улучшение наблюдалось для ситуации, когда для обучения использовалось не менее 5 образцов.

3. Вторая серия экспериментов

Для проверки качества распознавания речи использовался самостоятельно сформированный речевой корпус русской речи с записями произнесения чисел от одного до десяти. При записи сигналов использовались микрофон Acme CD-930 (полоса частот 30…16000 Гц; чувствительность -58 ± 2 дБ) и звуковой редактор Audacity [1]. Параметры записанных в формате wav сигналов: 44100 Гц; 32 бита; моно.

Первый этап. Исследовалась зависимость качества автоматического распознавания отдельных слов от следующих параметров:

- количество дикторов, обучающих систему (один либо два диктора);
- длительность паузы между произнесёнными, в режиме тестирования, словами (0,3 с или 1 с);
- варьирование словарем фонетических транскрипций слов (пробовались различные варианты транскрибирования смягченных согласных и ударных гласных).

Результаты экспериментов свидетельствуют:

- качество распознавания значительно (более 10%) понижается, если САРР обучать на образцах речи одного диктора, а тестировать на образцах речи другого диктора;
- качество распознавания снижается в меньшей степени (до 10%), если обучение САРР производить не одним, а двумя дикторами;
- количество ошибок распознавания типа «вставка» растет с увеличением длительности пауз между словами тестовой выборки;
- варьирование словарем фонетических транскрипций слов сказывается на качестве распознавания, в проведенных экспериментах степень такого влияния достигала 5%.

Второй этап. Исследовалась зависимость эффективности работы САРР от количества образцов слов в обучающей выборке, частоты дискретизации и транскрибирования шумов дыхания.

Результаты распознавания, приведенные в табл. 5, свидетельствуют, что рост объема обучающей выборки с 2-х до 20 образцов, приводит к повышению качества распознавания при условии, что частота дискретизации выбрана не слишком большой (16 кГц в данном эксперименте). Увеличение частоты дискретизации с 16 до 44 кГц существенно снижает качество распознавания, что поясняется ростом влияния фоновой шумовой помехи. В отличие от первой серии экспериментов, шумы дыхания не подавлялись путем «ручного вырезания», а распознавались системой НТК. С этой целью при транскрибировании обучающей выборки использовались специальные транскрипции vdoh (после стартовой транскрипции sil) и vydoh (перед финальной транскрипцией sil).

Третий этап. Исследовалась зависимость эффективности работы САРР от применения процедуры подавления шумов и транскрибирования шумов дыхания диктора. Подавление шумов осуществлялось с помощью соответствующего режима программы Audacity. Обучение системы производилось одним диктором, для каждого слова в обучающей выборке использовалось 20 образцов произнесения. Частота дискретизации составляла 16 кГц.

Результаты распознавания приведены в табл. 6. из которой следует, что рассмотренные способы подавления помех весьма эффективны. Приведенные данные хорошо согласуются с известными результатами исследований эффективности процедуры шумоподавления [6].

4. Обсуждение результатов

Обобщим результаты проведенных экспериментов.

Влияние уровня фонового шума. Результаты, полученные в обеих сериях экспериментов, свидетельствуют о принципиальной возможности значительного улучшения качества распознавания при использовании процедуры шумоподавления при формировании обучающей и тестовой выборок речевых сигналов. Отсюда следует вывод о перспективности исследований, направленных на поиск наиболее эффективных, с позиций задачи распознавания речи, алгоритмов шумоподавления.

Влияние шумов диктора. С ростом длительности пауз между распознаваемыми словами возрастает количество ошибок типа «вставка». Причиной тому наличие фоновых шумов, окружающих слова, а также наличие шумов диктора (шум дыхания, «щелчки» губ и т.п.), принимаемых САРР за речевые сигналы. Применение процедур шумоподавления и транскрибирования шумов диктора позволяет практически полностью избавиться от ошибок

типа «вставка». Другой, более сложный, подход к решению указанной проблемы состоит в нахождении границ речевых сигналов с помощью детекторов речевой активности (Voice Activity Detector).

Таблица 5.

Кол-во вар. произн.	Процент распознанных слов, %		
	Правильно	Замена	Вставка
	Без транскр. / С транскр.	Без транскр. / С транскр.	Без транскр. / С транскр.
	Частота дискретизации 44100 Гц		
2	40 / 60	60 / 40	50 / 10
5	40 / 90	60 / 10	0 / 0
10	40 / 90	60 / 10	0 / 0
20	30 / 80	70 / 20	0 / 0
	Частота дискретизации 16000 Гц		
2	50 / 50	50 / 50	20 / 0
5	90 / 100	10 / 0	0 / 0
10	90 / 100	10 / 0	10 / 0
20	100 / 100	0 / 0	0 / 0

Таблица 6.

SNR, дБ	Процент распознанных слов, %					
	Без транскрибирования			С транскрибированием		
	Правильно	Замена	Вставка	Правильно	Замена	Вставка
	Без процедуры подавления шумов					
0	-	-	-	-	-	-
10	26	74	3	52	48	6
20	67	33	3	81	19	3
45	93	7	3	96	4	0
	После процедуры подавления шумов					
6	-	-	-	16	84	20
25	54	46	4	61	39	4
40	82	18	1	93	7	4
60	79	21	1	95	5	0

Влияние типа параметризации. В отсутствие значительных фоновых шумов (при отношениях сигнал-шум 30 дБ и более) наилучшие результаты распознавания, в большинстве случаев, достигаются при использовании MFCC_0_D_A-параметризации. Однако при распознавании зашумленной речи (при отношениях сигнал-шум 20 дБ и меньше) увеличение мерности признакового пространства нецелесообразно, поскольку приводит к значительному (до 25%) ухудшению качества распознавания, по сравнению с наиболее простой MFCC_0-параметризацией.

Влияние словаря транскрипций на качество распознавания. «Ручная» модификация автоматически сформированных транскрипций позволяет, в принципе, добиться некоторого улучшения качества автоматического распознавания речи.

Однако, учитывая значительную трудоёмкость ручного редактирования транскрипций, нетрудно заключить, что такое редактирование уместно лишь при формировании небольших речевых корпусов. В случае больших речевых корпусов практически неизбежным является использование транскрипций, сгенерированных автоматически, на базе готовых словарей фонетических транскрипций [10].

Влияние количества обучающих дикторов на качество распознавания. Использование в обучающей выборке образцов произнесения слов, принадлежащих различным дикторам, приводит к снижению качества распознавания тестового сигнала, в котором присутствует речь лишь одного из обучающих дикторов. Причиной данного явления является несоответствие параметров обученной системы и параметров речевого сигнала тестовой выборки.

Влияние частоты дискретизации. Снижение частоты дискретизации с 44 кГц до 16 кГц положительно сказывается на качестве распознавания речи, что объясняется уменьшением мощности шумового процесса при практически неизменной мощности речевого сигнала.

Выводы

Эффективность функционирования САРР существенно зависит как от характера и уровня шумов, так и от параметров настройки САРР.

Проведенные в данной работе экспериментальные исследования позволили не только проверить справедливость предположений о возможных способах борьбы с помехами при распознавании речи, но и получить количественные оценки действенности таких способов. В частности, подтверждено предположение о целесообразности применения процедуры предварительной фильтрации фонового шума. Кроме того, показано, что с помощью специального транскрибирования возможно подавление негативного влияния импульсной помехи типа «щелчков» губ, нестационарной помехи типа шумов дыхания, и даже в некоторой степени нейтрализовать действие шумовой фоновой помехи, окружающей слова.

Чрезвычайно интересным и полезным для инженерных приложений представляется вывод о целесообразности оптимизации состава вектора кепстральных параметров, с учётом уровня фонового шума. Кроме того, полезным для инженерных приложений является вывод о допустимости автоматического транскрибирования в тех задачах, где ручное транскрибирование невозможно из-за временных и финансовых ограничений.

Литература

1. Audacity / [Электронный ресурс]. – Режим доступа к ресурсу: http://audacity.sourceforge.net. Дата обращения: 20.04.2012.
2. *Documentation* for HTK / [Электронный ресурс]. – Режим доступа к ресурсу: http://htk.eng.cam.ac.uk/docs/docs.shtml. Дата обращения: 3.11.2012.
3. *Furui S.* 50 years of progress in speech and speaker recognition / [Электронный ресурс]. – Режим доступа к ресурсу: http://www.furui.cs.titech.ac.jp/publication/2005/SPCOM05.pdf. Дата обращения: 3.11.2012.
4. *Nelson N.* HTK MFCC Study / [Электронный ресурс]. – Режим доступа к ресурсу: http://speech-research.com/mfccStudy.html Дата обращения: 13.12.2012.
5. *Sony* Sound Forge ™ 10 Pro / [Электронный ресурс]. – Режим доступа к ресурсу: http://www.sonycreativesoftware.com/. Дата обращения: 20.10.2012.
6. Verbout S.M. Signal Enhancement for Automatic Recognition of Noisy Speech / Master degree dissertation for the degree of Master of Science. – May, 1994. – 84 p.
7. *Краткий* учебный курс по HTK / [Электронный ресурс]. – Режим доступа к ресурсу: http://www.speech.com.ua/htk_course.html . Дата обращения: 10.12.2012.
8. *Продеус А.Н.,* Ладошко О.Н. Оптимизация алгоритмов системы распознавания

речи с использованием инструментария НТК. // Электроника и связь. – 2007. – № 4. – С. 53–60.
9. *Продеус А.Н.,* Литвинов С.В. Моделирование системы распознавания украинской речи с применением инструментария НТК. // Электроника и связь. – 2009. – № 1. – С. 88–94.
10. *Фонетическая* транскрипция. Современный русский язык / [Электронный ресурс]. – Режим доступа к ресурсу: http://morfema.ru/publ/15-1-0-3. Дата обращения: 9.05.2012.

УДК 621.3

Про вплив завад й параметрів налаштування на якість функціонування системи автоматичного розпізнавання мови

С.А. Йолкін, **А.Г. Ляшенко**, **В.П. Овсяник**, д-р техн. наук, **А.М. Продеус**, д-р техн. наук
Національний технічний університет України «Київський політехнічний інститут»,
пр. Перемоги, 37, Київ-56, 03056, Україна.

Розробку систем автоматичного розпізнавання мови доцільно робити з урахуванням впливу перешкод різної природи (шум акустичного оточення, реверберація, фільтрація й кодування мовного сигналу в системах зв'язку). В даній роботі експериментально досліджено вплив завад і деяких характеристик системи автоматичного розпізнавання мови на якість розпізнавання. Отримано рекомендації з оптимізації параметрів системи автоматичного розпізнавання мови для декількох сценаріїв її використання. Бібл.10, рис. 1, табл.6.

Ключові слова: *розпізнавання мови, завади, шумозаглушення, параметризація, транскрибування, мовленнєвий корпус.*

UDC 621.3

On noise and feature settings action on quality of the automatic speech recognition system

S.A. Yolkin, **A.G. Liashenko**, **V.P. Ovsyanik**, Dr.Sc., **A.N. Prodeus**, Dr.Sc.
National Technical University of Ukraine «Kyiv Polytechnic Institute»,
37 Prospect Peremogy, Kiev 03056, Ukraine.

The development of automatic speech recognition systems is expedient to make with the different nature of interference (noise acoustic environment, reverberation, filtering and encoding of speech in communication systems). The effect of noise and some features of the automatic speech recognition system on the recognition quality is experimentally investigated in this paper. Recommendations on the parameters optimization of the automatic speech recognition system for several scenarios of usage are obtained. Reference 10, figures 1, tables 6.

Keywords: *speech recognition, noise, noise reduction, parameterization, transcription, speech corpus.*

References

1. Audacity / [Online]. Available at: http://audacity.sourceforge.net. (Rus) (17.01.2013)
2. Documentation for HTK [Online]. Available at: http://htk.eng.cam.ac.uk/docs/docs.shtml. (17.01.2013).
3. Furui S. 50 years of progress in speech and speaker recognition [Online]. – Available at: http://www.furui.cs.titech.ac.jp/publication/2005/SPCOM05.pdf. (17.01.2013)
4. Nelson N. HTK MFCC Study. [Online]. Available at: http://speech-research.com/mfccStudy.html (17.01.2013).
5. Sony Sound Forge™ 10 Pro [Online]. Available at: http://www.sonycreativesoftware.com/.(17.01.2013)
6. Verbout S.M. Signal (1994), [Enhancement for Automatic Recognition of Noisy Speech]. Master degree dissertation for the degree of Master of Science. May, P.84.
7. A short course on HTK [Online]. Available at: http://www.speech.com.ua/htk_course.html. (Rus) (17.01.2013).
8. Prodeus A.N., Ladoshko O.N. (2007), [Optimization of speech recognition system algorithms with usage of HTK toolkit]. Electronics and Communication. no 4. pp. 53–60. (Rus)
9. Prodeus A.N., Litvinov S.V. (2009), [Modeling of Ukrainian speech recognition system with usage of HTK toolkit]. Electronics and Communication. no 1. pp. 88–94. (Rus)
10. Fonetic transcription. Modern Russian language [Online]. Available at: http://morfema.ru/publ/15-1-0-3. (Rus) (17.01.2013).

Поступила в редакцию 17 января 2013 г.