ИССЛЕДОВАНИЕ ХАРАКТЕРИСТИК ВОКАЛИЗОВАННЫХ ПАУЗ СПОНТАННОЙ УКРАИНСКОЙ РЕЧИ

О. Н. ЛАДОШКО

Национальный технический университет Украины ”КПИ”, Киев

Показано, что закономерности изменения траектории частоты основного тона можно использовать в качестве классификационных признаков при обнаружении вокализованных пауз в спонтанной речи. Полученные результаты могут быть использованы при построении детекторов вокализованных пауз для систем автоматического стенографирования.

ВВЕДЕНИЕ

В течение последних двух лет украинскими исследователями проводятся исследования по автоматическому распознаванию спонтанной украинской речи (АРСУР). В работе [1] были проведены эксперименты по автоматическому распознаванию (АР) спонтанной украинской речи (УР) в задаче автоматизированного стенографирования. Обнаружено, что главной проблемой в работе АРСУР является ухудшение показателя надёжности распознавания речи, обусловленное наличием особенностей в спонтанной речи (СР) (от 2,44% [1] до 4,9 % [2] всех слов), таких как вокализованные паузы (ВП) (паузы, заполненные звуками «э-э» и т.п.), невокализованные паузы (паузы, заполненные звуками дыхания, кашля и т.п.), суржик украинско-русской речи, редуцирование (искажение) звуков слов, растягивания, повторы и обрывы слов.

На основе учета особенностей СР и ручной коррекции стенограммы в [1, 3] представлены результаты очистки данных от спонтанных особенностей, которая позволила улучшить показатели надёжности распознавания речи в среднем от 1,25% до 6,45% для разных исследуемых выборок. В работе [2] был сформулирован подход к обобщению спонтанных особенностей украинской речи негативно влияющих на работу АРСУР и введена расширенная система их разметки, необходимая для исследования влияния нарушений спонтанной речи на показатели надёжности системы АРСУР.

Как было показано в [1, 2] наибольшую часть особенностей СР составляют ВП, которые представлены в СР в виде самостоятельных ВП, ограниченных паузами речи, а также растягиваниями звуков в словах (38,4% от всех особенностей СР). Предполагается, что для улучшения надёжности работы системы АР слитной речи в условиях распознавания СР, необходимо учитывать особенности СР, а в отдельных случаях ВП очищать СР от них.

В связи с этим в данной статье исследуется одна из наиболее распространенных особенностей СР – *вокализованные паузы* (ограниченные паузами речи), путём оценки траекторий частоты основного тона (ЧОТ) СР в условиях окружающего шума.

1. ВОКАЛИЗОВАННЫЕ ПАУЗЫ СПОНТАННОЙ РЕЧИ

Вокализованные паузы представляют собой паузы в речи говорящего, в продуцировании которых участвуют голосовые связки. К ним мы относим все возможные «акания», «экания», «мэкания» (см. рис 1). Произнося любую из ВП, диктор даёт понять собеседнику, что ему необходимо дополнительное время для формирования дальнейшей речи.

Не смотря на попытку выделить класс ВП, ограниченных паузами речи было обнаружено, что некоторые ВП представлены одновременным образованием двух тонов различной высоты при произнесении одного звука, именуемое эффектом *диплофонии* (см. рис 1 а, б, в). В ходе исследования было обнаружено, что некоторые ВП оканчиваются обрывом траектории ЧОТ вследствие нестабильности колебаний голосовых связок в конце фрагмента ВП (см. рис 1б, в). Такой эффект на участке ВП связан с появлением хрипоты в голосе. Предварительные исследования [4] базовых акустических характеристик показали, что произношение ВП связано с незначительным изменением положения артикуляционных органов, которое приводит к незначительным флуктуациям контура ЧОТ (в среднем в пределах 5,5 Гц).

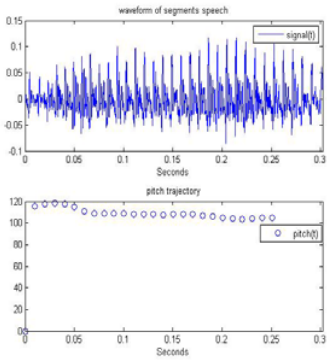

Рис. 1а. ВП типа «е»

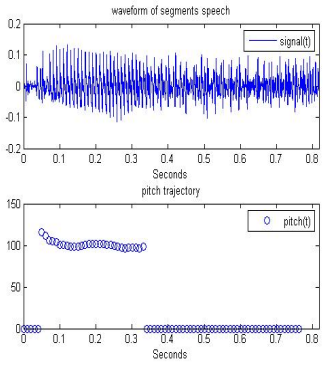

Рис. 1б. ВП типа «ем»

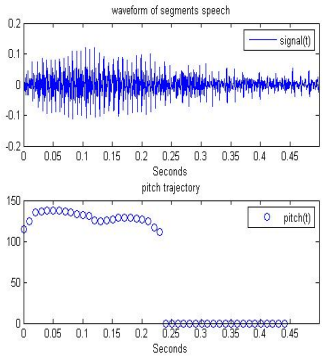

Рис. 1в. ВП типа «ео»

Рис. 1. Некоторые примеры вокализованных пауз (осциллограмма сигнала вверху) и полученные траектории частоты основного тона (ЧОТ – нижний график)

В [4] было предложено исследовать характеристики поведения ЧОТ для поиска признаков, позволяющих в дальнейшем вынести процедуру автоматического детектирования ВП на этап предварительной обработки. Это позволит не расширять словарь системы АР слитной речи при распознавании СР, а значит, и не увеличивать время обучения системы АРСУР.

2. МЕТОДЫ И СРЕДСТВА ПРОВЕДЕНИЯ ИССЛЕДОВАНИЙ

Для исследований использовались стенограммы заседаний Верховной Рады Украины [1, 3]. В тексты стенограмм была внесена ручная сегментация ВП, пауз речи и не ВП в отдельные сегменты, ограниченные временными метками [4]. Сегменты без ВП представляли собой сочетания слов и отдельно стоящие слова. Длительные паузы речи исключались из исследования. Упрощение ручной сегментации текста стенограммы обеспечила специальная программа для аннотирования и сегментации речевых сигналов – Transcriber 1.5.1 [5].

Для последующего анализа тексты стенограмм преобразовывались в необходимый формат данных для их последующего исследования. Автоматический анализ и формирование необходимых векторов исследуемых данных выполнялись при помощи специального разработанного автором *анализатора текстовых данных*, который стал необходимой частью автоматизации исследования речевых корпусов большого объёма.

В таблице 1 приведены автоматически извлеченные, с помощью анализатора, из текстов стенограмм статистические данные исследуемого речевого корпуса записей

заседаний Верховной Рады Украины [1, 3]. Общая длительность анализируемой выборки звукозаписей составила 4,4 часа речи.

Таблица 1. Статистические характеристики исследуемого материала

№ файла п/п	Кол-во слов, шт.	Общ. время, мин.	Общее кол-во ВП		ВП вид «е»		ВП вид «а»		ВП вид *еа*	Общ. кол. дикт., число
			шт.	%	шт.	%	шт.	%		
1	3734	29	98	2,6	82	2,2	5	0,1	17	13
2	4123	31	115	2,8	64	1,6	44	1,1	35	12
3	4839	35	274	5,7	222	4,6	40	0,8	23	16
4	5088	40	163	3,2	134	2,6	13	0,3	15	17
5	3345	28	63	1,9	53	1,6	3	0,1	11	14
6	4717	31	132	2,8	73	1,5	41	0,9	25	19
7	5101	41	180	3,5	152	3,0	7	0,1	18	17
8	3905	27	69	1,8	54	1,4	9	0,2	10	10
Сум.	34852	262	1094	3,1	834	2,4	162	0,5	154	118

Оценка траекторий ЧОТ СР в условиях окружающего шума была реализована в виде программы на основе предложенного в работе [6] помехоустойчивого выделителя ЧОТ. Общая структурная схема выделителя ЧОТ приведена на рис. 2:

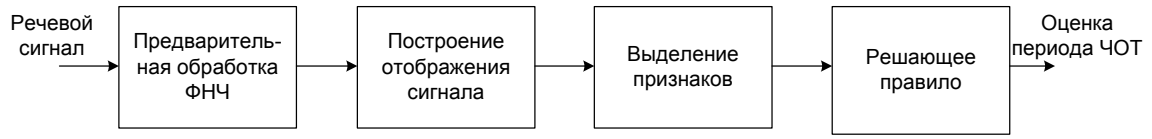

Рис. 2. Общая структурная схема выделителя ЧОТ

Предварительная обработка сигнала осуществляется при помощи ФНЧ с частотой среза 1,2 кГц (частота дискретизации сигналов 22050 Гц). Оценка ЧОТ проводиться в окне анализа длительностью 50мс с шагом 10мс. Построение отображения сигнала строилось для сдвигов p от $p_{\min} = 37$ до $p_{\max} = 550$ на основе расчета функции нормированной автокорреляции (ФНАК) $R(p)$ [6]:

$$R(p) = \frac{\sum_{n=1}^{N-p} x(n)x(n-p)}{\sqrt{\sum_{n=1}^{N-p} x^2(n)x^2(n-p)}} \qquad (1)$$

Известно [6], что оценка периода ЧОТ заключается в поиске *глобального максимума* в диапазоне $p_{\min} \le p \le p_{\max}$. Для реальных речевых сигналов задача поиска глобального максимума ФНАК сводиться к оценке *положительных локальных максимумов* ФНАК отдельного кадра, образующих набор возможных кандидатов $\{p_m\}$ с дальнейшим применением правил сортировки кандидатов в ЧОТ. Кроме того для зашумлённых сигналов необходимо учитывать информацию о возможной траектории ЧОТ для каждого кадра анализа на основе оценок ЧОТ смежных кадров [6, 7].

Таким образом, проводится совокупный анализ группы смежных кадров, учитывающий вероятности возможных кандидатов на оценку ЧОТ для каждого из кадров путём поиска пути, максимизирующего общую вероятность появления оценок ЧОТ для группы кадров с условием непрерывности траектории ЧОТ для гласных звуков [6].

Таблица 2. Сравнение результатов различных выделителей ЧОТ (значения в Гц)

ЧОТ, Гц	начало	конец	среднее	min	max	std
Praat	127,257	129,498	128,430	124,977	130,381	1,718
Реализован.выделитель ЧОТ	126,724	128,947	128,399	125,284	130,473	1,669
Разница, Гц	0,533	0,551	0.031	0,307	0,092	0,049

Выбор остальных параметров выделителя ЧОТ осуществлялся экспериментально в соответствии с идеями [6, 8] путём минимизации ошибок и выбором допустимой задержки выдачи оценки ЧОТ при работе выделителя ЧОТ. Основным ограничением анализа ЧОТ ВП была длительность анализируемого сегмента речи, которая должна была быть не менее 0.1 секунды.

Проверка правильности получаемых оценок была проведена путём сравнения результатов работы [9] данного метода и алгоритма реализованного в программном продукте Praat [10]. Сравнение оценок ЧОТ проводились на одном протяженном гласном звуке (табл.2), а также на случайных выборках, данных путём визуального сопоставления траекторий ЧОТ в ходе проведения исследований. Схема измерений ЧОТ представлена на рис. 3.

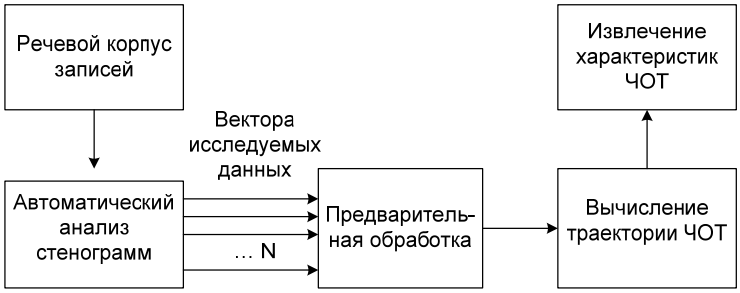

Рис. 3. Схема измерений ЧОТ

3. РЕЗУЛЬТАТЫ ИССЛЕДОВАНИЙ

Для данных из таб. 1 экспериментально были получены все значения траекторий ЧОТ в исследуемом контексте.

Изменение траектории ЧОТ через ВП аппроксимировалось пиковыми значениями [11] F0 в окружении ВП и значениями F0 на границах ВП, усреднённые значения которых приведены в таб.3.

Детально были исследованы два файла (1 и 6, табл. 3). В ходе исследований получены характеристики изменения траектории ЧОТ для различных реализаций ВП (103 и 125 – для 1 и 6 файлов соответственно) для каждого диктора в отдельности. Некоторые зависимости поведения траекторий ЧОТ для отдельных дикторов приведены на рис. 4.

Выявлено, что траектория ЧОТ на границе речь-ВП-речь убывает на участке ВП и возрастает вне этого промежутка в большинстве рассмотренных случаев (на 103 ВП – 88,3% случаев для файла 1 и на 125 ВП – 80,8% случаев для 6 файла).

Таблица 3. Усреднённые значения F0 на ВП и окружающих пиках F0 для 8 файлов

файл	Пред. Пик F0, Гц	Нач. F0, Гц	Конец, F0, Гц	Посл. пикF0, Гц
1	208,8	174,1	167,8	216,4
2	180,4	143,4	130,8	189,8
3	185,6	150,8	143,8	194,3
4	198,0	158,6	153,2	195,4
5	234,3	166,9	167,4	195,5
6	181,0	151,8	139,9	180,4
7	189,0	139,8	135,8	185,8
8	157,0	125,3	120,5	169,6

Возрастание на границе ВП - следующий пик F0 не происходит лишь в 7,8% и 15,2% случаев. Среднее значение разницы частот на этой границе составляет – 41 Гц и 36,8 Гц для файлов 1 и 6 соответственно. Убывание ЧОТ на границе предыдущий пик F0 – ВП не происходит в 3,9% и 4% случаев соответственно для рассматриваемых файлов. Среднее значение разницы частот на этой границе составляет – 47,3 Гц и 45,9 Гц для файлов 1 и 6 соответственно.

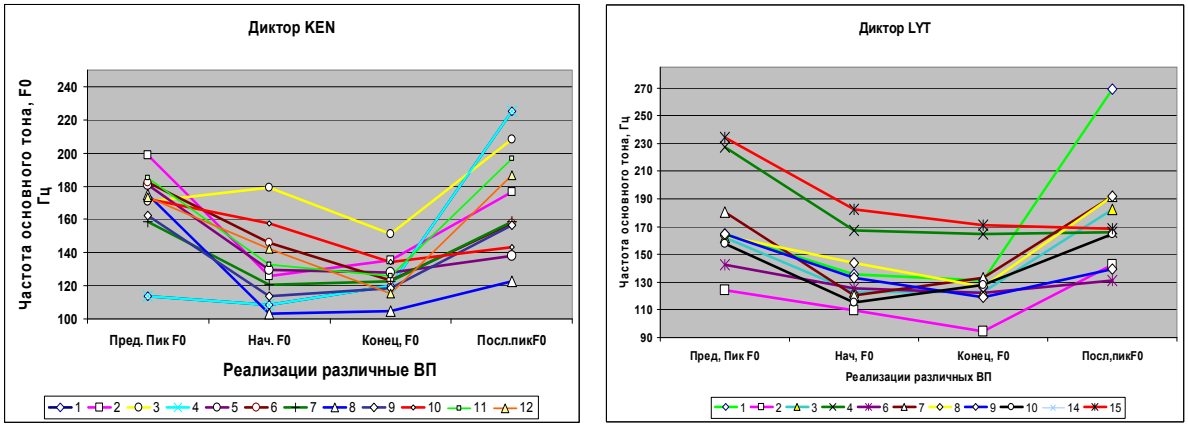

Рис. 4. Поведение траекторий ЧОТ различных реализаций ВП для нескольких дикторов.

На промежутке начало-ВП – конец-ВП наблюдается убывание траектории ЧОТ в 64,1% случаев (103 ВП) и в 70,4% случаев (125 ВП). Были проведены исследования траекторий ЧОТ в отсутствии ВП. Было обнаружено, что изменение ЧОТ на границах в переделах произнесения фразы без ВП значительно изменяется лишь на границе пауз речи.

Отсюда следует вывод, что такое поведение траектории ЧОТ СР УР является признаком, который позволяет обнаружить часть ВП в СР путём исследования характеристик траекторий ЧОТ в условиях шума.

ЗАКЛЮЧЕНИЕ

Представлены результаты исследования одной из наиболее распространённых особенностей спонтанной речи – вокализованных пауз. Приведены автоматически извлеченные из текстов стенограмм статистические данные ВП исследуемого речевого корпуса записей заседаний Верховной Рады Украины. Путём оценки траекторий частоты

основного тона (ЧОТ) СР в условиях окружающего шума получены траектории ЧОТ исследуемого речевого корпуса.

Экспериментально выявлены закономерности движения контура ЧОТ на участках контуров ЧОТ до ВП и после. Изменение траектории ЧОТ аппроксимировалось пиковыми значениями F0 в окружении ВП и значениями F0 на границах ВП. Выявлено, что траектория ЧОТ на границе речь-ВП-речь убывает на участке ВП в большинстве рассмотренных случаев (на 103 ВП – 88,3% случаев). На промежутке начало-ВП – конец-ВП наблюдается убывание траектории ЧОТ в 64,1% случаев (103 ВП). Полученные результаты могут быть использованы при проектировании систем автоматического стенографирования.

ЛИТЕРАТУРА

1. Ладошко О.Н., Пилипенко В.В. Аннотация и учет речевых сбоев в задаче автоматического распознавания спонтанной украинской речи // Искусственный интеллект. – Донецк – № 3. – 2010. – С. 238-248.
2. Ladoshko O. N., Prodeus A. N. Annotation of Ukrainian Spontaneous Speech. Proceedings of XXXI International Scientific Conference Electronics and Nanotechnology. 12-14 April, 2011, Kyiv, Ukraine.
3. Ладошко О.Н., Пилипенко В.В. Аннотация и учет речевых сбоев в задаче автоматического распознавания спонтанной украинской речи // Международная научно-техническая конференция «Искусственный интеллект. Интеллектуальные системы ИИ-2010», Тези доп., Том 1 – Донецк, 2010. – С. 223-227.
4. Ладошко О.Н. Исследование акустических особенностей вокализованных пауз спонтанной речи // Международная научно-техническая конференция «Искусственный интеллект. Интеллектуальные системы ИИ-2011», Тези доп., Том 3 – Донецк, 2011. – С. (в печати).
5. Barras C., Geoffrois E., Wu Z., Liberman M. Transcriber: development and use of a tool for assisting speech corpora production // Speech Communication special issue on Speech Annotation and Corpus Tools, Vol 33, No 1-2, January 2000.
6. Бабкин В. В. Помехоустойчивый выделитель основного тона речи. // 7-я Международная Конференция и Выставка Цифровая Обработка Сигналов и её Применение DSPA-2005. – Москва 16-18 марта. – 2005г.– С.175-178.
7. Баронин С. П. Автокорреляционный метод выделения основного тона речи. Пятьдесят лет спустя // Речевые технологии. – М. – №2. – 2008. – С. 3-12.
8. Adam P. Vogel, Paul Maruff, Peter J. Snyder, James C, Mundt. Standartization of pitch-range settings in voice acoustic analysis. // Behavior Research Methods. – 41(2)/ – 2009. – pp. 318-324.
9. Gerhard D. Pitch extraction and fundamental frequency: history and current techniques // Technical report TR-CS 2003–06. 2003. University of Regina, Saskatchewan, Canada.
10. P. Boersma and D. Weenink, "Praat, a system for doing phonetics by computer", Institute of Phonetic Sciences of the University of Amsterdam", pp. 132-182, 1999.
11. E.E. Shriberg. Phonetic consequences of speech disfluency // ICPhS99, Speech Technology and Research Laboratory SRI International, Menlo Park, 1999.