УДК 004.394.2

# ИССЛЕДОВАНИЕ ВЛИЯНИЯ ХАРАКТЕРИСТИК ТЕЛЕФОННОГО КАНАЛА СВЯЗИ НА НАДЕЖНОСТЬ РАСПОЗНАВАНИЯ ФОНЕМ

**Ладошко О.Н.**
Национальный технический университет Украины ”Киевский политехнический институт”
кафедра акустики и акустоэлектроники
ladoshko@gmail.com

***Аннотация***

***Ладошко О.Н. Исследование влияния характеристик телефонного канала связи на надёжность распознавания фонем.*** *Проведены исследования влияния характеристик телефонного канала связи на точность распознавания контекстно-независимых (монофонов) и контекстно-зависимых (трифонов) фонем слитной речи в зависимости от вида параметризации речевого сигнала. Получены оценки точности распознавания фонем, при различных условиях записи обучающей и тестовой выборок.*

**Общая постановка проблемы**

В работе [1], где изложены результаты экспериментов по автоматическому распознаванию спонтанной украинской речи, было показано, что спонтанная речь требует предварительной обработки с учетом таких особенностей как всевозможные вставки слов, вокализованные паузы и др. для последующего её распознавания системой автоматического распознавания слитной речи. Однако кроме очистки спонтанной речи от её особенностей, перед исследователями также ставиться задача робастного распознавания спонтанной речи в условиях искажений, связанных с наличием канала передачи (channel distortion).Ухудшение надежности распознавания в таком случае объясняется различием характеристик каналов записи обучающей выборки и тестовой выборки. Известно, что проблемы, связанные с различными условиями записи речевых сигналов являются наиболее критичным ограничением современных систем распознавания речи [2]. К таким условиям в первую очередь относят наличие передаточной характеристики коммуникационного канала связи.

В данной работе исследуется влияние характеристик телефонного канала связи на точность распознавания фонем. Для построения акустических моделей контекстно-зависимых (монофонов) и контекстно-независимых фонем (трифонов) использовались скрытые Марковские модели (HMM – hidden Markov models). Распознавание проводилось для дикторонезависимого режима работы системы автоматического распознавания фонем слитной речи. Исследования проводились при MFCC и PLP параметризации речевых сигналов на речевой базе, записанной в лабораторных условиях с высоким качеством, и базе этих же речевых сигналов, но пропущенных через телефонный канал.

**Методы и средства проведения исследований**

Исследование робастности системы автоматического распознавания фонем в отношении всевозможных мультипликационных помех и искажений телефонного канала связи в данной работе заключается в поиске признаков, инвариантных к значительному изменению условий записи обучающей и тестовой выборок. К таким условиям относятся:

наличие канала связи (transmission channels) – дальнего или локального действия;

наличие устройств, преобразующих сигнал, например, микрофоны, телефонные трубки, гарнитуры.

Известно, что основное требование к параметризации сигнала (признакам, извлекаемых из речевого сигнала) при дикторонезависимом распознавании речи заключается в том, чтобы при этом сглаживались индивидуальные особенности голосов дикторов.

Предполагают, что речевой сигнал стационарен на промежутках времени порядка нескольких миллисекунд. В ходе анализа речевой сигнал разбивается на блоки данных (окна). На основе данных, полученных путём взвешивания речевого сигнала окном, вычисляются вектора признаков.

В данной работе исследуется два широко распространенных способов параметризации речевого сигнала, а именно мэл-частнотные кепстральные коэффициенты – MFCC (Mel-frequency cepstral coefficients) и перцепционные коэффициенты линейного предсказания – PLP (perceptual linear predictive) [4]. Коэффициенты MFCC и PLP представляют собой некоторую разновидность кепстра, что позволяет говорить [3, 4, 5] об их эффективности при работе в условиях мультипликативных шумов. Процедура получения MFCC коэффициентов на практике состоит в следующем: выборку значений кепстра $c_n$ вычисляют через выборку значений

$$M_j = \sum_{i=0}^{N/2} m_j c_n , \quad j = 0,1,..., P , \tag{1}$$

полученных путем усреднения непараметрической оценки спектра треугольными весовыми функциями (рис.1):

$$c_n = \sqrt{\tfrac{2}{N}} \sum_{j=1}^{N} m_j \cos(\tfrac{\pi \cdot n}{N}(j - 0.5)) \tag{2}$$

Ширина весовых функций постоянна на нелинейной мел-шкале частот. За счет использования мэл-шкалы удается учесть нелинейную зависимость слухового восприятия от частоты речевого сигнала:

$$Mel(f) = 2595 \log(1 + f/700) \tag{3}$$

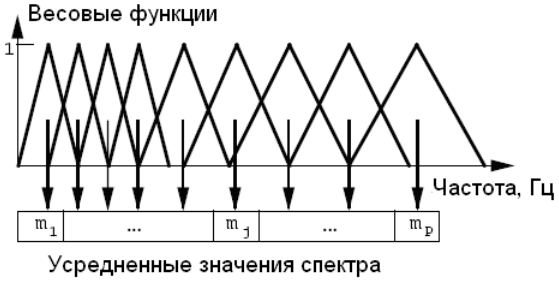


Рисунок 1 - Мел-шкала и усредняющие треугольные функции

Таким образом, производиться формальное сглаживание спектра речевого сигнала, которое в свою очередь значительно упрощает моделирование речи за счет снижения размерности вектора признаков. Необходимость использовать оценку спектра, получаемую при помощи быстрого преобразования Фурье (БПФ), приводит к тому, что процесс получения коэффициентов MFCC в вычислительном смысле является более затратным. Поэтому на практике используют другой подход к вычислению коэффициентов MFCC: выборку значений кепстра $c_n$ вычисляют через коэффициенты $\alpha_k$, $k = 0 \ldots K$, параметрической (авторегрессионной) оценки спектра речевого сигнала, с помощью рекуррентного соотношения [4]:

$$c_n = -\alpha_n - \frac{1}{n} \sum_{i=1}^{n-1} (n - i) \alpha_i c_{n-i} \tag{4}$$

Коэффициенты $\alpha_k$, $k = 0 \ldots K$ при этом находят путем решения системы уравнений Юла-Уолкера [5]:

$$r_n = -\sum_{k=1}^{M} \alpha_k r_{n-k} \quad \text{для} \quad 1 \le n \le M + 1 \quad \text{и} \quad \sigma^2 = r_0 + \sum_{k=1}^{M} \alpha_k r_{-k} \quad \text{для} \quad n = 0 \tag{5}$$

где $r_n$ – выборки оценки корреляционной функции сегмента речевого сигнала; $\sigma^2$ – оценка дисперсии белого шума, воздействующего на авторегрессионный фильтр, характеризуемый коэффициентами $a_k$, $k = 0 \ldots K$. Одним из достоинств коэффициентов MFCC является их независимость, что в свою очередь позволяет моделировать функции плотности вероятности с помощью диагональной ковариационной матрицы.

Альтернативой использованию MFCC коэффициентов являются коэффициенты перцепционного линейного предсказания PLP (perceptual linear predictive) [3]. Техника использования PLP параметризации основана на психоакустических концепциях при оценивании спектра:

спектральный анализ в критических полосах частот;

кривые равной громкости;

нелинейная связь между интенсивностью и воспринимаемой громкостью звука.

Извлечение PLP коэффициентов основано на стандартном мэл-частотном анализе спектра Фурье с помощью гребенки фильтров (рис. 1). Спектр Фурье предварительно вычисляется по $N$ – отсчетам сигнала $s_1, \ldots, s_N$. Коэффициенты, полученные на выходе гребенки фильтров $M_j = \sum_{i=0}^{N/2} m_j c_i$, $j = 0,1,\ldots,K$ взвешиваются кривой равной громкости, которая задана эмпирически в виде:

$$E(\omega_j) = \frac{\left(\omega^2 + 1200^2\right) \cdot \omega^4}{\left(\omega^2 + 400^2\right)^2 \cdot \left(\omega^2 + 3100^2\right)} \tag{6}$$

где $a_j$ – частота $j$-го треугольного окна (рис. 1) $M_j^{'} = M_j E(\omega_j)$ и затем сжимаются путём извлечения кубического корня $M_j^{''} = \sqrt[3]{M_j^{'}}$. Далее путём расчета обратного преобразования Фурье на основе значений $M_j^{''}$ вычисляют коэффициенты линейного предсказания LP (linear predictive) по второму методу, описанному выше.

Базовая система распознавания фонем моделировалась с помощью программного инструментария HTK (Hidden Markov Model (HMM) Toolkit – инструментарий на основе Скрытых Марковских Моделей) [6] и с учетом рекомендаций работ [7, 8]. Использовались лево-правые HMM модели, состоящие из 3-х состояний без пропуска с непрерывными гауссовыми смесями (CDHMM – continuous density HMM). Анализ речевого сигнала проводиться с помощью окна Хэмминга длительностью 25 мс, с шагом анализа 10 мс. Данное окно применяется для каждого кадра речи перед дальнейшей обработкой. Речевой сигнал пропускается через фильтр высоких частот с передаточной характеристикой $P(z) = 1 - 0{,}97z^{-1}$. Количество треугольных окон для проведения анализа на нелинейной мел-шкале частот равно 26. Вычислялись 12 кепстральных коэффициентов, дополненные логарифмом энергии. С целью учета изменения параметров во времени коэффициенты кепстра, и логарифм энергии были дополнены первой (префикс _D) и второй производными (префикс _A) [4]. К PLP коэффициентам, вместо дополнения логарифмом энергии, к вектору параметров добавлялся нулевой кепстральный коэффициент (префикс _0). Путем добавления префикса _Z, проводилась нормализация кепстрального среднего (CMN – cepstral mean normalization) [4]. Данная операция позволяет устранить различные эффекты, связанные с искажениями частотных характеристик записывающих устройств или каналов передачи, путём вычитания среднего значения, вычисленного за длительный интервал, из последовательности кепстральных коэффициентов.

Обучение акустических моделей начиналось с плоского старта (flat start), при этом создавалась универсальная унимодальная модель (гауссиан). Прототипы создаваемых моделей содержали одну гауссову смесь с одним потоком. На дальнейших циклах обучения постепенно увеличивалось количество гауссовых смесей до максимального значения для

монофонов и трифонов равное 20. Количество прямых и обратных ходов при увеличении количества смесей составило 4. Монофоны полученные на стадии обучения с одной гауссовой смесью использовались для клонирования контекстно-зависимых фонем – трифонов. После одного цикла обучения трифоны связывались посредством алгоритма кластеризации, при построении дерева решений на основании тренировочных данных, с учетом правил английского языка. Эти правила заключались в генерировании вопросов о левом и правом контексте, используя установленные классы фонем.

**Результаты экспериментов**

Эксперименты по определению точности распознавания фонем $Accuracy = \frac{N - S - D - I}{N} \times 100\%$, где $N$ – относительное количество правильно распознанных эталонных меток, $S$ – количество замен, $D$ – количество удалений, $I$ – количество вставок [6], проводились на материале речевых корпусов TIMIT и NTIMIT. TIMIT – это речевой корпус, содержащий свыше 5 часов цифровых звукозаписей различных английских фраз, произнесённых 630 дикторами на 8 диалектах американского английского.

Все звукозаписи имеют временную фонемную разметку, выполненную профессиональными фонетистами. Речевой корпус разбит на два непересекающихся множества: обучающее и тестовое [9]. Речевой корпус NTIMIT построен на основе речевого корпуса TIMIT. Звукозаписи речевого корпуса TIMIT были пропущены через телефонные каналы американской телефонной компании NYNEX и заново оцифрованы. Это позволило представить в речевом корпусе NTIMIT звукозаписи с искажениями, характерными для естественного телефонного канала связи [10]. Тестирование проводилось на подмножестве Core Test Set исследуемых баз TIMIT и NTIMIT. Данное подмножество состояло из 24 дикторов, 2 мужчин и одной женщины из региона с присущим ему диалектом (8 диалектов). Каждый диктор читал различные предложения. Таким образом, подмножество Core Test Set содержало 192 предложения для каждого диктора.

Результаты экспериментов приведены в таблицах 1–4, из которых видно, что точность распознавания фонем существенно ухудшается при пропускании речевого сигнала через телефонный канал связи. Точность распознавания фонем повышается при использовании трифонов и PLP- параметризации сигнала. Максимальные значения точности достигаются при различных значениях гауссовых смесей. Тем не менее, использование трифонов, вместе с PLP параметризацией, основанной на психоакустических концепциях, не повышают точность распознавания до уровня точности распознавания фонем при совпадении условий записи тестовой и обучающей выборок.

Таблица 1 – Точность распознавания монофонов при MFCC_E_D_A_Z параметризации

| Обучение | Тест | mix4 | mix6 | mix8 | mix10 | mix12 | mix14 | mix16 | mix18 | mix20 |
|---|---|---|---|---|---|---|---|---|---|---|
| TIMIT | TIMIT | 55,6 | 58,0 | 59,3 | 59,9 | 60,9 | 61,0 | 61,6 | 61,4 | 61,8 |
| NTIMIT | NTIMIT | 38,6 | 41,2 | 42,3 | 44,1 | 45,2 | 45,5 | 45,8 | 46,1 | 46,4 |
| TIMIT | NTIMIT | 21,5 | 22,2 | 22,7 | 22,8 | 22,9 | 23,6 | 24,2 | 24,3 | 24,7 |

Таблица 2 – Точность распознавания трифонов при MFCC_E_D_A_Z параметризации

| Обучение | Тест | mix4 | mix6 | mix8 | mix10 | mix12 | mix14 | nmix16 | mix18 | mix20 |
|---|---|---|---|---|---|---|---|---|---|---|
| TIMIT | TIMIT | 62,9 | 63,8 | 64,0 | 63,8 | 63,7 | 63,1 | 62,7 | 62,0 | 62,1 |
| NTIMIT | NTIMIT | 45,8 | 47,4 | 47,7 | 47,7 | 47,7 | 47,6 | 47,8 | 47,3 | 46,6 |
| TIMIT | NTIMIT | 25,8 | 25,5 | 24,5 | 24,4 | 23,9 | 24,5 | 23,5 | 23,4 | 23,1 |

Таблица 3 – Точность распознавания монофонов при PLP_0_D_A_Z параметризации

| Обучение | Тест | mix4 | mix6 | mix8 | mix10 | mix12 | mix14 | nmix16 | mix18 | mix20 |
|---|---|---|---|---|---|---|---|---|---|---|
| TIMIT | TIMIT | 55,7 | 57,4 | 59,1 | 60,2 | 60,5 | 60,7 | 61,2 | 61,7 | 61,9 |
| NTIMIT | NTIMIT | 40,1 | 41,8 | 42,7 | 44,0 | 44,5 | 45,2 | 45,7 | 46,2 | 47,2 |
| TIMIT | NTIMIT | 22,9 | 23,2 | 23,0 | 22,9 | 23,1 | 23,6 | 23,6 | 23,4 | 23,5 |

Таблица 4 – Точность распознавания трифонов при PLP_0_D_A_Z параметризации

| Обучение | Тест | mix4 | mix6 | mix8 | mix10 | mix12 | mix14 | nmix16 | mix18 | mix20 |
|---|---|---|---|---|---|---|---|---|---|---|
| TIMIT | TIMIT | 62,6 | 63,6 | 63,9 | 63,5 | 62,8 | 62,8 | 62,7 | 62,6 | 61,9 |
| NTIMIT | NTIMIT | 46,3 | 47,1 | 47,5 | 47,7 | 47,1 | 46,8 | 47,0 | 47,2 | 46,0 |
| TIMIT | NTIMIT | 26,9 | 26,5 | 27,0 | 26,6 | 26,9 | 26,2 | 25,5 | 25,4 | 25,1 |

Работа с NTIMIT приводит к заметно худшим результатам, по сравнению с TIMIT, что можно пояснить потерей информации из-за таких особенностей телефонных линий связи, как ограниченность полосы частот и неравномерность АЧХ тракта, составляющая около 11 дБ в полосе частот от 300 до 3400 Гц (рис. 2). При этом обеспечивается высокая степень разборчивости речи, хорошая естественность её звучания и создаются большие возможности для вторичного уплотнения телефонных каналов.

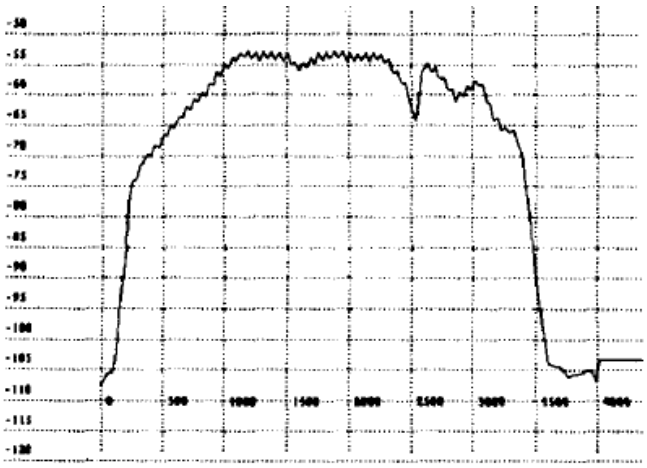

Рисунок 2 –. Амплитудно-частотная характеристика канала передачи

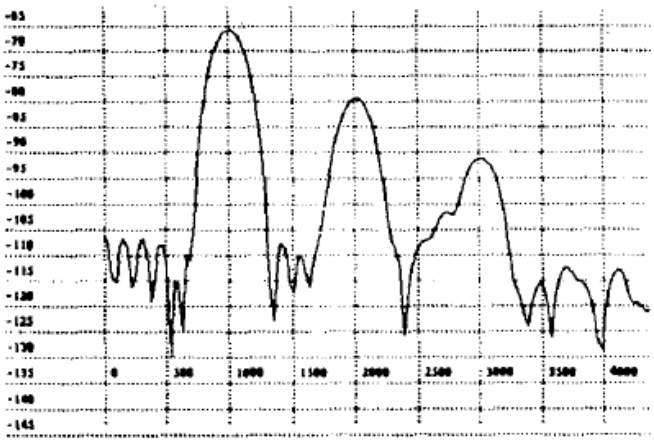

Рисунок 3 –. Отклик сигнала 1 кГц тон из локального центрального офиса

Для опорного сигнала частотой 1 кГц в телефонном тракте наблюдаются нелинейные искажения, проявляющиеся в виде возникновения высших гармоник (рис. 3). Степень нелинейности ФЧХ телефонного тракта NYNEX в работе [10] не указана.

Таким образом, из результатов таблиц 1-4 можно сделать вывод о необходимости проведения обучения акустических моделей при распознавании телефонной речи на выборке, записанной в аналогичных условиях. Выполнение этого условия позволяет обеспечить точность распознавания не менее 46-47 %.

Качество распознавания речи, взятой из базы TIMIT, системой распознавания, обученной на базе NTIMIT, в данной работе не исследовалось, поскольку такая ситуация при практическом использовании представляется маловероятной.

**Выводы**

Качество распознавания фонем существенно ухудшается при пропускании речевого сигнала через телефонный канал связи. И хотя использование трифонов и PLP-

параметризации сигнала позволяет улучшить качество распознавания, тем не менее, это улучшение недостаточно для достижения уровня, соответствующего ситуации одинаковых условий записи тестовой и обучающей выборок. Получены оценки степени ухудшения качества распознавания фонем слитной речи, обусловленного влиянием таких характеристик телефонного канала связи как ограниченная полоса частот, неравномерность АЧХ тракта, а также нелинейные искажения.

Продемонстрирована необходимость проведения обучения акустических моделей при распознавании телефонной речи на выборке, записанной в аналогичных условиях. Выполнение этого условия позволяет обеспечить точность распознавания не менее 46-47 %.

## Список литературы

1. Ладошко О.Н., Пилипенко В.В. Аннотация и учет речевых сбоев в задаче автоматического распознавания спонтанной украинской речи [Текст] / О. Н. Ладошко, В. В. Пилипенко // Искусственный интеллект. – Донецк – № 3. – 2010. – С. 238-248.

2. Rabiner L. R. Applications of Voice Processing to telecommunications // Proceedings of the IEEE, 82, pp. 199, February 1994.

3. Hermansky H. Perceptual linear predictive (PLP) analysis of speech // J. Acoust. Soc. Am. 111 – 1990.–Vol.87, №4, pp. 1738–1752.

4. Picone, J.W. Signal modeling techniques in speech recognition // Proceedings of the IEEE, 81, pp. 1215, September 1993.

5. Rabiner L. Fundamentals of Speech Recognition. // Prentice-Hall International Inc. – 1993. – 507 p.

6. Young S., Everman G. Moore, J. Odell, D. Ollason, V. Valtchev, Woodland P. The HTK Book // Cambridge University Engineering Department. – 2005, pp. 354.

7. HTK training for TIMIT from Cantab Research [Electronic resourse] / Интернет-ресурс. - Режим доступа: www/ URL: http://www.cantabResearch.com/HTKtimit.html - Multiple Choices.

8. Ладошко О.Н., Продеус А.Н. Оптимизация алгоритмов системы распознавания речи с использованием инструментария HTK [Текст] / О.Н. Ладошко, А.Н. Продеус // Электроника и связь – 2007. – № 4(39). – С. 53–60.

9. Zue V., Seneff S., Glass J. Speech database development at MIT: TIMIT and beyond // Speech Communication. – 1990. – Vol. 9, № 4. – P.351-356.

10. Jankowski C., Kalyanswamy A., Basson S., Spitz J. NTIMIT: A Phonetically Balanced, Continuous Speech, Telephone Bandwidth Speech Database // Proc. of ICASSP-90. – 1990. – P. 109-112.