ИССЛЕДОВАНИЕ ВЛИЯНИЯ ПАРАМЕТРИЗАЦИИ РЕЧЕВОГО СИГНАЛА И ХАРАКТЕРИСТИК КАНАЛОВ СВЯЗИ НА НАДЕЖНОСТЬ АВТОМАТИЧЕСКОГО РАСПОЗНАВАНИЯ ФОНЕМ

О. Н. ЛАДОШКО

Национальный технический университет Украины "КПИ", Киев

Экспериментально оценено влияние характеристик телефонных каналов связи и способов параметризации речевых сигналов на точность распознавания фонем.

ВВЕДЕНИЕ

Задача робастного распознавания спонтанной речи в условиях различия характеристик каналов записи обучающей и тестовой выборок весьма актуальна [1, 2] К таковым относятся различие передаточных характеристик каналов связи, различие микрофонов, различие расстояний от рта до микрофона.

В данной работе исследовалось влияние характеристик телефонного канала связи на точность распознавания фонем. Для построения акустических моделей контекстно-независимых фонем (трифонов) использовались скрытые Марковские модели (HMM – Hhidden Markov models). Распознавание проводилось для дикторонезависимого режима работы системы автоматического распознавания фонем слитной речи. Исследования проводились при MFCC и PLP параметризации речевых сигналов. Для обучения использовалась речевая база, записанная с высоким качеством (отношение сигнал-шум не менее 40 дБ). Для распознавания использовались различные виды искаженной речи: 1) естественная речь на выходе одноканального телефонного канала связи; 2) синтетическая речь на выходе телефонного канала связи, сформированная с использованием системы «искусственный рот».

1. МЕТОДЫ И СРЕДСТВА ПРОВЕДЕНИЯ ИССЛЕДОВАНИЙ

Исследование робастности систем автоматического распознавания (САР) фонем заключается в поиске признаков, инвариантных к возможному значительному различию «образцовой» речи, использовавшейся при обучении САР, и условий реальной речи на выходе телефонных каналов связи.

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 [3]. 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 [4, 5, 6, 7] об их эффективности при работе в условиях мультипликативных шумов. Процедура получения MFCC коэффициентов на практике состоит в следующем: выборку значений кепстра 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:

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

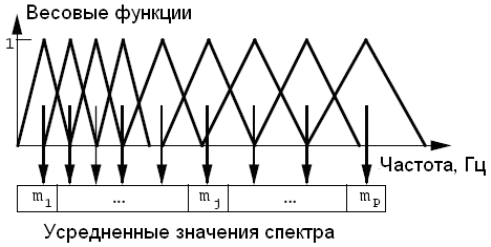

Рис.1. Мел-шкала и усредняющие треугольные функции

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 α_k, $k = 0 \ldots K$, параметрической (авторегрессионной) оценки спектра речевого сигнала, с помощью рекуррентного соотношения [4]:

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

Одним из достоинств коэффициентов MFCC является их статистическая независимость, что в свою очередь позволяет моделировать функции плотности вероятности с помощью диагональной ковариационной матрицы [7].

Альтернативой использованию MFCC коэффициентов являются коэффициенты перцепционного линейного предсказания PLP (perceptual linear predictive) [4, 7]. 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 $s_1,...,s_N$. Коэффициенты, полученные на выходе гребенки фильтров $M_j = \sum_{i=0}^{N/2} m_j c_i$, $j = 0,1,...,K$ взвешиваются кривой равной громкости, которая задана эмпирически в виде:

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

где ω_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).

В данной работе базовая система распознавания фонем моделировалась с помощью программного инструментария HTK (Hidden Markov Model (HMM) Toolkit – инструментарий на основе Скрытых Марковских Моделей) [8] и с учетом рекомендаций работ [9, 10]. Использовались лево-правые HMM модели, состоящие из 3-х состояний без пропуска с непрерывными гауссовыми смесями. Анализ речевого сигнала проводился с помощью окна Хэмминга длительностью 25 мс, с шагом анализа 10 мс. Данное окно применялось для каждого кадра речи перед дальнейшей обработкой. Речевой сигнал пропускался 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) [5, 9]. 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), направленная на подавление эффектов, обусловленных различием частотных характеристик каналов записи и передач речевых сигналов [9].

Обучение акустических моделей начиналось с плоского старта (flat start) [9], 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, равного 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. При построении дерева решений на основании тренировочных данных, после одного цикла обучения трифоны связывались посредством алгоритма кластеризации, с учетом правил английского языка.

2. РЕЗУЛЬТАТЫ ИССЛЕДОВАНИЙ

Эксперименты по определению точности распознавания фонем, характеризуемой мерой $Accuracy = \frac{N - S - D - I}{N} \times 100\%$, где N – относительное количество правильно распознанных эталонных меток, S – количество замен, D – количество удалений, I – количество вставок [8], проводились на материалах речевых корпусов TIMIT, NTIMIT и STC-TIMIT [10-12]. Данные по результатам распознавания трифонов на корпусах TIMIT-NTIMIT из таблиц 1 и 2 приведены для сравнения из работы [2].

Заметим, что речевые корпуса NTIMIT и STC-TIMIT получены из речевого корпуса TIMIT путем пропускания его речевого материала через различные телефонные каналы американской телефонной компании NYNEX. В результате речевой корпус NTIMIT содержит звукозаписи с искажениями, характерными для естественного телефонного канала связи [11], а корпус STC-TIMIT получен путем пропускания звуковых сигналов корпуса NTIMIT через коммутатор одноканального телефонного канала передачи [13].

Тестирование проводилось на подмножестве Core Test Set всех исследуемых баз. Данное подмножество соответствовало группе из 24 дикторов, из регионов с присущим им диалектом (8 диалектов). Каждый диктор читал различные предложения. Таким образом, подмножество Core Test Set содержало 192 предложения для каждого диктора. Эксперименты проводились на корпусе данных STC-TIMIT при одинаковых условиях тестовой и обучающей выборок.

Результаты экспериментов приведены в таблицах 1–3, из которых видно, что точность распознавания фонем корпуса STC-TIMIT существенно ухудшается при пропускании речевого сигнала через телефонный канал связи [2]. Точность распознавания фонем повышается при использовании трифонов, вместе с PLP параметризацией, основанной на психоакустических концепциях.

В работе [2] было показано, что работа с NTIMIT приводит к заметно худшим результатам, по сравнению с TIMIT. Результаты работы с корпусом STC-TIMIT также свидетельствуют об ухудшении, по сравнению с результатами на TIMIT, но являются лучшими, по сравнению с NTIMIT (см. табл.1, 2, 3). Это можно объяснить тем, что искажения в STC-TIMIT обусловлены влиянием канала передачи коммутатора и телефонной линии локального действия [13] без мультипликативных искажений, внесенных телефонными трубками, которые представленные в NTIMIT [2, 12].

Потери надежности 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 (NTIMIT) [2].

Действительно, базы TIMIT и NTIMIT содержат сигналы, дискретизированные с частотой 16 кГц, при эффективной полосе пропускания 6,4 кГц [2]. Между тем, полоса частот 3,4…6,4 кГц не содержит речевой информации [12]. На стадии предобработки треугольные фильтры покрывают весь частотный диапазон от нуля вплоть до частоты Найквиста. Чтобы убрать нежелательные компоненты в полосе от 3,4…6,4 кГц из дальнейшего анализа, заданное число каналов гребёнки фильтров было распределено равномерно по мел-шкале от края до края результирующей полосы пропускания от 300 до 3400 Гц.

Таблица 1. 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
STC-TIMIT	STC-TIMIT	44,6	45,8	46,8	47,3	48,1	47,7	47,6	47,4	47,2

Таблица 2. 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
STC-TIMIT	STC-TIMIT	45,7	47,4	48,5	49,1	49,2	49,1	48,8	48,6	48,3

Таблица 3. Точность распознавания трифонов при MFCC_E_D_A параметризации и ограничении полосы пропускания от 300 до 3400Гц

Обучение	Тест	mix4	mix6	mix8	mix10	mix12	mix14	mix16	mix18	mix20
TIMIT	TIMIT	57,4	58,8	59,1	59,7	59,7	59,4	58,4	58,6	58,2
NTIMIT	NTIMIT	46,8	47,5	47,8	47,7	48,0	47,1	46,4	45,9	45,8
NTIMIT_Z	NTIMIT_Z	46,5	47,6	47,9	48,8	48,6	48,2	47,9	47,6	48,0

Из таблицы 3 видно, что полосовая фильтрация позволяет повысить надежность распознавания даже при использовании стандартного подхода на основе MFCC коэффициентов. Незначительное улучшение результатов распознавания может быть объяснено тем, что в случае NTIMIT вся неинформативная часть сигнала выше 4 кГц эффективно отфильтрована самим телефонным каналом. Кроме того существенное улучшение результатов наблюдается во всех случаях применения вычитания среднего значения, вычисленного за длительный интервал, из последовательности кепстральных коэффициентов (см. префикс _Z).

Таким образом, на основании результатов, представленных в таблицах 1-3, можно сделать вывод о целесообразности подавления неинформативных частотных компонентов. Кроме того, результаты распознавания, полученные на различных телефонных речевых корпусах, таких как NTIMIT и STC-TIMIT, свидетельствуют о необходимости учета влияния передаточных характеристик всех устройств, входящих в канал связи.

ЗАКЛЮЧЕНИЕ

Качество распознавания фонем для речевого корпуса STC-TIMIT существенно уступает таковому для речевого корпуса TIMIT. И хотя использование трифонов и PLP-параметризации сигнала позволяет улучшить качество распознавания, тем не менее, это улучшение недостаточно для достижения уровня, соответствующего ситуации распознавания чистой речи TIMIT при одинаковых условиях записи тестовой и обучающей выборок. Получены оценки степени ухудшения качества распознавания фонем слитной речи, обусловленного влиянием таких характеристик телефонных каналов связи как ограниченная полоса частот, неравномерность АЧХ тракта, а также нелинейные искажения.

Продемонстрирована необходимость подавления неинформативных частотных компонентов, а также нормализации кепстрального среднего при распознавании телефонной речи на выборке, записанной в аналогичных условиях. Выполнение этих условий позволяет повысить точность распознавания до 47-49%.

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. *Ладошко О. Н.* Исследование влияния характеристик телефонного канала связи на надёжность распознавания фонем // III Международная научно-техническая конференция студентов, аспирантов и молодых ученых. Информационные управляющие системы и компьютерный мониторинг (ИУС и КМ-2012), Секция: Искусственный интеллект, нейросетевые и эволюционные алгоритмы, экспертные системы. 16-18 апреля, 2012г., Донецк – 2012.
3. *Rabiner L. R.* Applications of Voice Processing to telecommunications // Proceedings of the IEEE, 82, pp. 199, February 1994.
4. *Hermansky H.* Perceptual linear predictive (PLP) analysis of speech // J. Acoust. Soc. Am. 111 – 1990.–Vol.87, №4, pp. 1738–1752.
5. *Picone, J.W.* Signal modeling techniques in speech recognition // Proceedings of the IEEE, 81, pp. 1215, September 1993.
6. *Rabiner L.* Fundamentals of Speech Recognition. // Prentice-Hall International Inc. – 1993. – 507 p.
7. *Kacur J., Rozinaj G.* Building accurate and robust HMM models for practical ASR systems // Telecommunication Systems, Springer. – 2011. – Vol. 52, № 3 – P. 1683-1696.
8. *Young S., Everman G. Moore, J. Odell, D. Ollason, V. Valtchev, Woodland P.* The HTK Book // Cambridge University Engineering Department. – 2005, pp. 354.
9. *HTK* training for TIMIT from Cantab Research [Electronic resourse] / Интернет-ресурс. - Режим доступа: www/ URL: http://www.cantabResearch.com/HTKtimit.html - Multiple Choices.
10. *Ладошко О.Н., Продеус А.Н.* Оптимизация алгоритмов системы распознавания речи с использованием инструментария НТК // Электроника и *связь – 2007. – № 4(39). – С. 53–60.*
11. Zue V., Seneff S., Glass J. Speech database development at MIT: TIMIT and beyond // Speech Communication. – 1990. – Vol. 9, № 4. – P.351-356.
12. *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.
13. *Morales N., Javier T., Javier G., Colas J., Toledano D.T.* STC-TIMIT: Generation of Single-channel Telephone Corpus // Proc. of LREC-2008 – 2008. – P. 391-395.