ОЦЕНКА ЭФФЕКТИВНОСТИ АКУСТИЧЕСКИХ МОДЕЛЕЙ ДЛЯ СИСТЕМ РАСПОЗНАВАНИЯ РЕЧИ НА ОГРАНИЧЕННОМ МАТЕРИАЛЕ

Холявин Павел Андреевич (p.kholyavin@mail.ru)

СПбГУ, Санкт-Петербург, Россия

Abstract: The paper presents an analysis of acoustic models for speech recognition that are trained on limited data. Two kinds of training material were used: phonetically balanced texts and randomly chosen fragments of read speech, each recording was about 3 min long. Test fragments consisted of recordings of both spontaneous and read speech. The main factor that influences the model's performance is found out to be the amount of training material measured in number of sounds.

Key words: automatic speech recognition, phonetics, low-resourced ASR, acoustic modeling, phonetically balanced text

Введение

Одной из наиболее важных проблем в современном распознавании речи является поиск качественного обучающего материала для создания акустической модели, т. е. модели, описывающей фонетические характеристики звуков языка [Huang, Acero, Hon, 2001]. Запись больших речевых корпусов, снабжённых пословными транскрипциями или качественной сегментацией, требует больших временных и финансовых затрат. Кроме того, в последнее время особое развитие получает автоматическое распознавание речи на языках, количество речевого материала для которых ограничено [Besacier et al, 2014]. Существует несколько путей решения этой проблемы. Одним из них является использование в качестве обучающего материала записи телепередач, сопровождаемых точными субтитрами [Jaeyun Jang, Hauptmann, 1999]. Тем не менее, точным субтитрам часто предпочитаются более краткие, передающие общий смысл, что неприемлемо для обучения акустических моделей. Другим способом является сегментация только небольшого фрагмента крупного корпуса (10 минут) и последующая автоматическая разметка остальной части корпуса [Lamel, Gauvain, Adda, 2002]. Такой способ позволяет существенно упростить создание акустических моделей. Тем не менее, запись корпуса все еще представляет собой проблему. Таким образом, перспективным направлением исследований представляется поиск речевого материала, который по своему наполнению обеспечит более высокое качество распознавания, чем выбранный случайным образом из корпуса материал того же объема. Это позволит обеспечить создание корпусов, направленных на обучение акустических моделей, которые будут меньше по объему и, соответственно, потребуют меньших финансовых затрат, чем корпуса общего назначения, обучение с помощью которых даст схожие результаты.

В данной работе исследуется эффективность акустических моделей (АМ), обученных на ограниченном особым образом устроенном материале. В качестве такого материала был взят фонетически представительный текст «Был тихий серый вечер», разработанный на кафедре фонетики СПбГУ для задач, связанных с синтезом речи по тексту [Stepanova, 1988]. Под фонетически представительным текстом (ФПТ) понимается текст, в котором встречаются все частотные звукосочетания языка. Длина ФПТ «Был тихий серый вечер» составляет 2301 фонемоупотребление (данные основаны на идеальной транскрипции), и чтение этого текста занимает в среднем 3–4 минуты. Таким образом, ожидается, что в ФПТ обязательно встретятся единицы, вероятность появления которых в случайно выбранных 3 минутах текста невысока, и, соответственно, эффективность акустических моделей, обученных на ФПТ будет более высокой, чем эффективность моделей, обученных на отрывке текста длины, сравнимой с длиной ФПТ. Кроме того, дополнительной задачей исследования является исследование оптимальных правил транскрипции, используемых в произносительном словаре системы, и, как следствие, набора аллофонов, которым оперирует акустическая модель.

Методы и материалы

Материал в данной работе был взят из трех корпусов русской речи, созданных на кафедре фонетики СПбГУ: Intas – корпус, в котором спонтанная речь и чтение совпадают лексически [Bondarko et al, 2003], CORPRES – корпус профессионального чтения [Skrelin et al, 2010] и CoRuSS – корпус спонтанной речи [Kachkovskaia et al, 2016]. Из каждого корпуса были взяты записи фонетически представительного текста, на которых обучались акустические модели, а также отрывки чтения разной длины для обучения и чтение и спонтанная речь для оценки работы системы распознавания речи. Особенность корпуса Intas позволила более корректно

сравнить качество распознавания чтения и спонтанной речи.

Для проверки поставленной гипотезы был проведен ряд экспериментов. В каждом из них обучалась акустическая модель, а затем ее эффективность проверялась при помощи распознавания тестового фрагмента (чтение или спонтанная речь), при этом использовалась языковая модель, построенная на распознаваемом тексте: это позволило свести ошибки, связанные с несовершенством языковой модели, к минимуму; таким образом, качество распознавания практически целиком зависело от акустической модели. Эффективность оценивалась при помощи параметра WER (Word Error Rate).

Все эксперименты были проведены на основе инструментария распознавания речи Kaldi [Povey et al, 2011]. Использовался стандартный рецепт. Акустические модели были основаны на гауссовских смесях и скрытых марковских моделях (GMM-HMM) [Zolfaghari, Robinson, 1996], [Rabiner, 1989]. В качестве акустических признаков, извлекаемых из сигнала, использовались мел-частотные кепстральные коэффициенты (MFCC) [Mermelstein, 1976]. Триграммные языковые модели в формате ARPA обучались при помощи пакета SRI Language Modeling Toolkit [Stolcke, 2002] и конвертировались в конечные преобразователи, используемые Kaldi, с помощью библиотеки OpenFST [Allauzen et al, 2007]. При составлении произносительного словаря использовалась фонемная транскрипция по Ленинградской фонологической школе, кроме того, безударные и ударные гласные были обозначены отдельными знаками.

Результаты

В первой серии экспериментов сравнивалась эффективность акустической модели, обученной на 4 записях ФПТ из корпуса CORPRES, и модели, обученной на 4 отрывках чтения той же длины тех же дикторов. В качестве тестового фрагмента был взят отрывок спонтанной речи из корпуса CoRuSS. Акустическая модель, обученная на ФПТ, дала показатель WER, равный 57.94 %, в то время как модель, основанная на чтении — 55.18 %. Таким образом, можно сделать предварительный вывод, что ФПТ не дает преимущества по сравнению с корпусами чтения.

Во второй серии экспериментов сравнивалась эффективность акустической модели, основанной на 4 записях ФПТ из корпуса INTAS, и моделей, основанных на отрывках чтения тех же дикторов разной длины. В качестве тестовых фрагментов были взяты отрывки чтения и спонтанной речи тех же дикторов. Такая конфигурация позволила смоделировать дикторозависимую систему и, таким образом, снизить параметры WER для более адекватного сравнения.

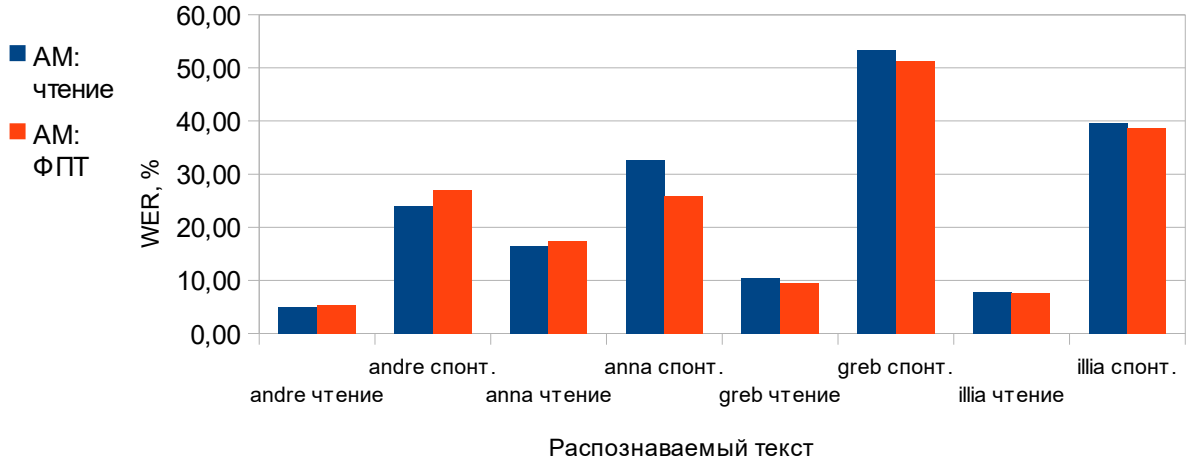

Рис. 1. Зависимость WER от акустической модели распознаваемого фрагмента

На рисунке 1 представлена зависимость качества распознавания речи от используемой акустической модели и характера распознаваемой речи: название каждого распознаваемого фрагмента состоит из идентификатора диктора и указания на тип материала — чтение или спонтанная речь. Длина каждого фрагмента совпадала с длиной записи ФПТ того же диктора. Из диаграммы видно, что, во-первых, чтение распознается лучше спонтанной речи, что не противоречит известным данным [Furui, 2019], во-вторых, эффективность ФПТ и чтения практически не отличается. Самое большое различие наблюдается при распознавании спонтанной речи диктора anna; это объясняется тем, что отрывок чтения этого диктора, входящий в обучающую выборку, хотя по длине и совпадал с записью ФПТ, оказался меньше по длине в фонемоупотреблениях. Таким образом, речь диктора anna была менее представлена в обучающем материале, чем речь всех остальных дикторов. В связи с этим дальнейшее измерение длины текстов проводилось преимущественно в фонемоупотреблениях.

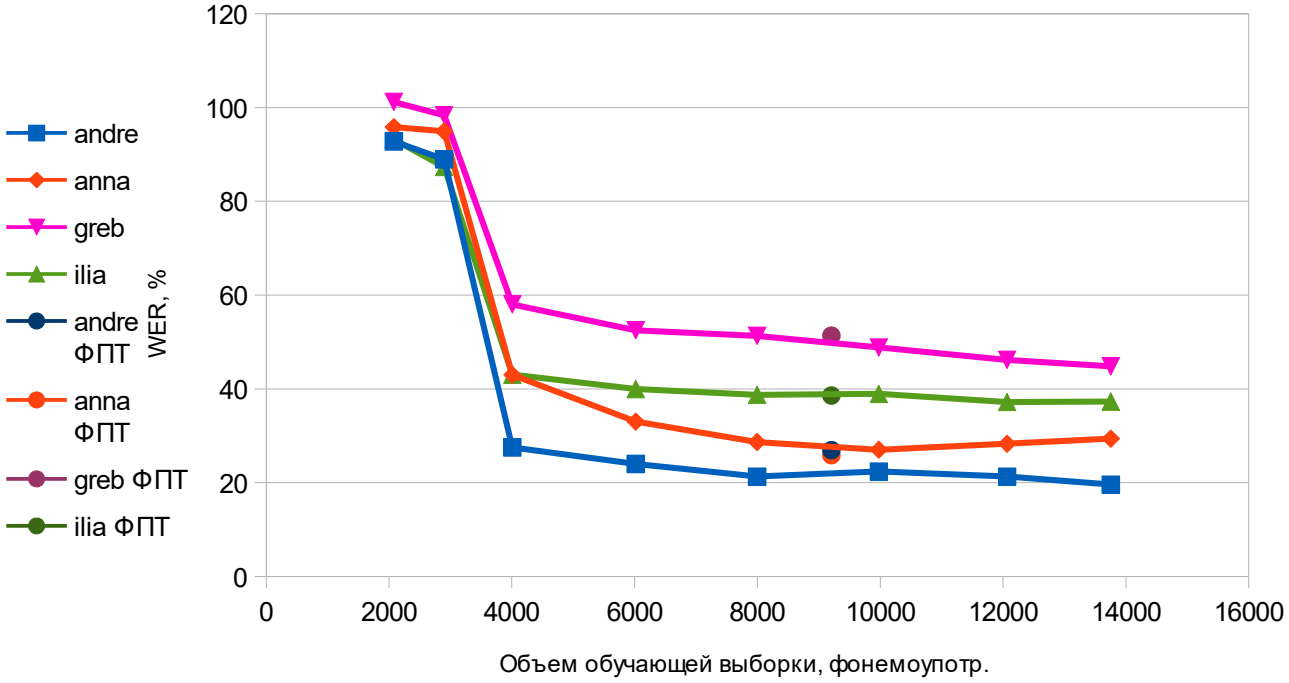

Рис. 2. Зависимость качества распознавания спонтанной речи от общего объема обучающей выборки, состоящей из речи четырех дикторов

На рисунке 2 приведена зависимость качества распознавания спонтанной речи от общего объема обучающей выборки для всех дикторов. Отдельными точками обозначены результаты для выборки из 4 записей ФПТ (см. рис. 1). Заметно, что при достижении объемом выборки некоторого порогового значения (3500-4000 фонемоупотреблений) происходит резкое падение WER, после чего параметр продолжает плавно уменьшаться. Видно также, что практически во всех случаях результаты, сравнимые с результатами для АМ, основанной на ФПТ, достигаются при меньшем объеме.

На рисунке 3 показана аналогичная зависимость для распознавания чтения. Видно, что при меньших значениях параметра WER пороговое значение, тем не менее, близко к значению из предыдущего эксперимента.

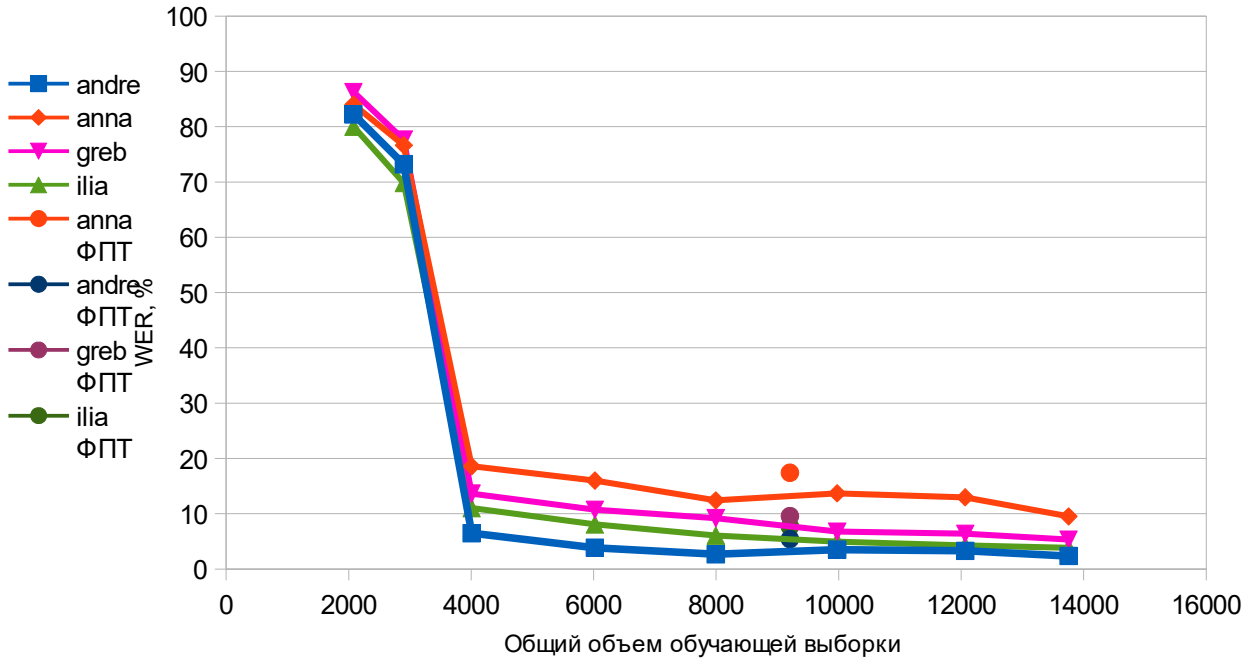

Рис. 3. Зависимость качества распознавания чтения от общего объема обучающей выборки, состоящей из речи четырех дикторов

В третьем эксперименте исследовались оптимальные правила транскрипции для произносительного словаря. В большей части экспериментов использовалась фонемная транскрипция с выделением ударения. В данном эксперименте в качестве обучающей выборки были взяты 4 записи ФПТ из корпуса CoRuSS, в качестве тестовой — запись спонтанной речи другого диктора из того же корпуса. При использовании стандартной транскрипции WER составил 57.94 %. Далее в список аллофонов был добавлен аллофон фонемы /а/ второй степени редукции [ъ]. Такое изменение дало улучшение параметра WER до 56.40 %. Добавление же в список аллофонов всех гласных после мягких согласных дало улучшение WER только до 56.27 %. Результаты эксперимента с более наглядным описанием правил транскрипции для каждого случая приведены в таблице 1.

Таблица 1. Зависимость качества распознавания речи от правил транскрипции произносительного словаря

Правило транскрипции	**WER, %**
Фонемная с выделением ударения Пример: «кораблями» [k a r a b l' a0 m' i]	57.94 %
То же с аллофоном [ъ] Пример: «кораблями» [k a2 r a b l' a0 m' i]	56.40 %
То же с [ъ] и аллофонами гласных после мягких согласных Пример: «кораблями» [k a2 r a b l' ia0 m' i]	56.27 %

Можно сделать вывод, что информация о позиционных аллофонах более значима для системы, чем информация о комбинаторных. Это связано с тем, что акустические модели в данном эксперименте были построены на трифонах, т. е. единицах, содержащих информацию о соседних звуках с обеих сторон. В целом, данный аспект исследования требует дальнейшей разработки. Вероятно, потребуется создание транскрипций, качественно отличающихся от словарных [Skrelin, 1999].

Выводы

Итак, из проведенной работы можно сделать следующие выводы:

1. Фонетически представительный текст, несмотря на особое статистическое устройство, не дает преимущества при использовании его записей для обучения акустичесих моделей в современных системах распознавания речи;

2. Минимальный необходимый объем обучающей выборки одинаков при распознавании как чтения, так и спонтанной речи и лежит в данном эксперименте в промежутке от 3000 до 4000 фонемоупотреблений;

3. Поиск наилучших правил транскрипции для произносительного словаря представляет собой перспективное направление исследований.

Список литературы:

Allauzen C., Riley M., Schalkwyk J., Skut W., Mohri M. (2007), OpenFst: A general and efficient weighted finite-state transducer library, International Conference on Implementation and Application of Automata, Springer, pp. 11–23

Bondarko L.V., Volskaya N.B., Tananaiko S.O., Vasilieva L.A. (2003), Phonetic Properties of Russian Spontaneous Speech, 15th ICPhS, Barcelona

Furui S. (2019), Recent advances in spontaneous speech recognition and understanding, IEEE Workshop on Spont. Speech Proc. and Rec., Tokyo, pp. 1–6

Huang X., Acero A., Hon H.-W. (2001), Spoken Language Processing: A Guide to Theory, Algorithm, and System Development, Prentice Hall PTR, Upper Saddle River, NJ, USA

Jaeyun Jang P., Hauptmann A. (1999), Improving acoustic models with captioned multimedia speech, Proceedings IEEE International Conference on Multimedia Computing and Systems, Florence

Kachkovskaia T., Kocharov D., Skrelin P., Volskaya N. (2016), CoRuSS – a New Prosodically Annotated Corpus of Russian Spontaneous Speech, LREC

Lamel L., Gauvain J.-L., Adda G. (2002), Lightly supervised and unsupervised acoustic model training, Computer Speech & Language, vol. 16. N 1, pp. 115–129

Mermelstein P. (1976), Distance measures for speech recognition, psychological and instrumental, Pattern recognition and artificial intelligence, vol. 116. pp. 374–388

Povey D., Ghoshal A., Boulianne G., Burget L., Glembek O. (2011), The Kaldi speech recognition toolkit, IEEE Signal Processing Society

Rabiner L.R. (1989), A tutorial on hidden Markov models and selected applications in speech recognition, Proceedings of the IEEE, vol. 77. N 2. pp. 257–286

Skrelin P.A. (1999), Segmentation and Transcription [Segmentatsiya i transkriptsiya], Izd-vo SPbGU, SPb

Skrelin P., Volskaya N., Kocharov D., Evgrafova K., Glotova O., Evdokimova V. (2010), CORPRES, International Conference on Text, Speech and Dialogue, Springer, pp. 392–399

Stepanova S.B. (1988) Phonetic properties of Russian speech: realization and transcription [Foneticheskiye svoystva russkoy rechi: realizatsiya i transkriptsiya], LGU, Leningrad

Zolfaghari P., Robinson T. (1996), Formant analysis using mixtures of Gaussians, Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP'96, pp. 1229–1232