ИНФОРМАЦИОННО-ТЕЛЕКОММУНИКАЦИОННЫЕ ТЕХНОЛОГИИ

УДК 551.74

ПРОБЛЕМЫ АВТОМАТИЧЕСКОГО РАСПОЗНАВАНИЯ СЛИТНОЙ РЕЧИ. МЕТОДЫ ОБРАБОТКИ ИСХОДНОГО РЕЧЕВОГО СИГНАЛА

THE PROBLEMS OF CONTINUOUS SPEECH AUTOMATIC RECOGNITION. THE METHODS OF PROCESSING THE ORIGINAL SPEECH SIGNAL

Н.И. Червяков, Н.Н. Кучукова
N.I. Chervyakov, N.N. Kuchukova

Кафедра прикладной математики и математического моделирования,
ФГАОУ ВО «Северо-Кавказский федеральный университет», 355009, Россия, Ставрополь, ул. Пушкина, 1

FSAEI HE "North-Caucasus Federal University", Department Mathematics and Mathematical Modelling,
1, Pushkin Street, Russia, Stavropol 355009

e-mail: k-fmf-primath@stavsu.ru, knn.storage@yandex.ru

Аннотация. Статья посвящена вопросам автоматического распознавания слитной речи. Проведен обзор современного состояния технологий распознавания, их применение в области транслирования и перевода речи в режиме реального времени. Рассматривается структура системы распознавания речи, включающая предварительную обработку речевого сигнала, акустическое моделирование, языковое моделирование и комбинирование. Особое внимание уделено этапу предварительной обработки сигналов, включающего выделение признаков речевого сигнала и их преобразование, в том числе выделение шумоустойчивых, адаптивных и дискриминативных характеристик. Приведены принципы построения таких акустических моделей, для которых применяются скрытые модели Маркова, оценка максимального правдоподобия, дискриминативное обучение.

Resume. This paper is about the problems of continuous speech automatic recognition. We conducted a review of the current state of recognition technologies, their application in area of speech broadcasting and translating in real time. We considered the structure of speech recognition system, including front-end processing, acoustic modeling, language modeling and system combination. Special attention is paid to front-end processing, including speech feature extraction and transformation, in particular noise robust, speaker-adaptive and discriminative features. We presented principles of acoustic modeling building, for which applied hidden Markov models, maximum likelihood estimation, discriminative training, speaker adaptation, noise adaptation, deep neural networks.

Ключевые слова: распознавание речи, предварительная обработка сигнала, акустическое моделирование, речевые характеристики, дискриминативное обучение

Keywords: speech recognition, processing the speech signal, acoustic modeling, speech features, discriminative training

Одной из самых острых проблем современной науки является создание надежной системы распознавания речи с высокой степенью устойчивости к шумам и искажениям, а также малым процентом ошибок. Модернизация аппаратной составляющей привела к появлению новых алгоритмов, использующихся в системах распознавания речи. Однако в данной области имеется большой ряд проблем, требующих детального изучения и решения. В данной статье приводится обзор имеющихся на настоящий момент технологий и методик распознавания слитной речи.

Современные системы распознавания слитной речи (LVCSR) включают несколько подсистем (см.рис. 1), каждая из которых содержит широкий набор задач и технологий их решения.

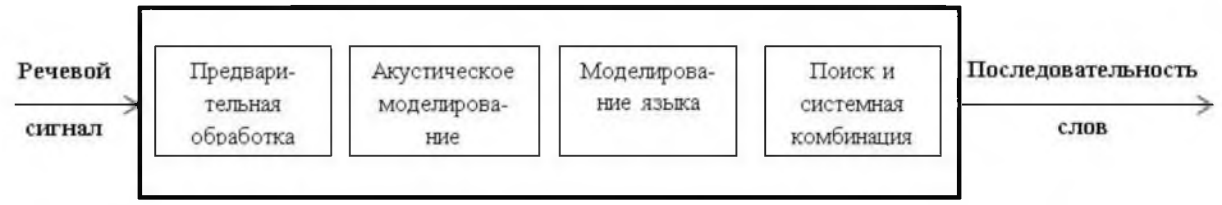

Рис. 1. Компоненты системы LVSR
Fig.1. Components of LVSR system

1. Предварительная обработка

Рассмотрим некоторые методы предварительной обработки, которые включают выделение признаков и преобразование, надежную обработку шумовых характеристик, а также оценку адаптивных и дискриминантных свойств, приведенных на рисунке 2.

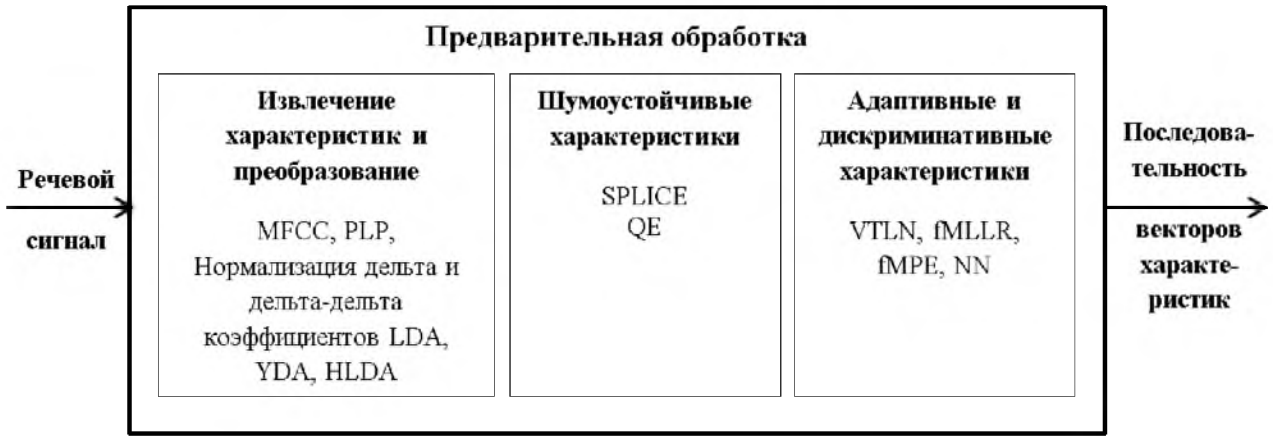

Рис. 2. Методы предварительной обработки речевого сигнала
Fig.2. Methods of front-end processing

1.1. Извлечение характеристик и их преобразование

Роль модуля предварительной обработки заключается в извлечении последовательности векторов X акустических характеристик из речевого сигнала S. В настоящее время это осуществляется посредством применения краткосрочного быстрого преобразования Фурье (FFT) речевого сигнала в течение 25 мс с временным окном в 100 р/с. Энергия соседних частот в пределах каждого кадра хранится вместе посредством мел-частотных фильтров, в которых ширина и расстояние между фильтрами определяются в ходе человеческой слуховой обработки. Далее к выходам фильтров применяется логарифм, и логарифмические тоновые спектры декоррелируется с помощью дискретного косинусного преобразования, преобразуясь в 13-мерный вектор кепстральных коэффициентов тоновой частоты (MFCC). В последнее время коэффициенты MFCC были заменены более шумоустойчивые представления, основанными на коэффициентах линейного перцепционного предсказания (PLP) [Hermansky, 1990].

В контексте распознавания слитной речи выделение признаков стало применяться с появлением двух важных технологий. Во-первых, использование основанных на дикторе среднего значения и дисперсии нормализации кепстральных коэффициентов. Тогда как метод оценки высказывания на основе вычитания кепстрального среднего (CMS) – хорошо известная технология, метод нормализации кепстральной дисперсии (CVN) на уровне диктора был разработан не так давно, в течение HUB-5 (или коммутаторной) эволюции [Chen et al, 2006]. Вторая идея заключается во включении временного контекста в кепстральные кадры. Общая практика заключается в вычислении скорости и коэффициентов ускорения (также называемых дельта и дельта-дельта коэффициентов) для соседних кадров внутри окна, обычно, из +/- 4кадров. Эти коэффициенты добавляются к статической кепстре для формирования окончательного вектора признаков [Furui, 1986].Эта специальная эвристика в современных LVCSR была заменена линейной матрицей проекции, отображающей вектор, полученный путем объединения последовательных кадров в пространстве низкой размерности. Проекция разработана таким образом, чтобы максимально отделить фонетические классы в трансформированном пространстве. Разделители обычно вычисляются с помощью критерия линейного дискриминантного анализа (LDA) [Saon et al, 2000]. Что-

бы сделать гипотезу моделирования диагональной ковариации более допустимой, пространственные признаки LDA «переворачиваются» посредством преобразования полу-привязанной ковариации (STC) [Gales, 1998], целью которой является сведение к минимуму вероятности потери между полной и диагональной ковариацией Гаусса. Использование каскада преобразований LDA и STC приводит к относительному уменьшению (на 10-15%) неправильно распознанных слов (WER) среди простых временных производных в некоторых задачах LVCSR.

1.2. Шумоустойчивые характеристики

В речевых сигналах часто содержатся внешние шумы, которые могут негативно повлиять на эффективность распознавания. Разработка шумоустойчивых методов имеет решающее значение для обеспечения надежности распознавания речи. Один из таких алгоритмов, называемый SPLICE [Deng, 2000], что означает «стерео кусочно-линейный компенсатор внешних воздействий», был предложен для распознавания речи в условиях нестационарного шума. Сущность данного алгоритма заключается в улучшении характеристик путем замены помех в зашумленном речевом сигнале наиболее вероятным вектором коррекции, который является ожидаемой разницей между чистой и зашумленной речью, ассоциируемой с наиболее вероятной областью акустического пространства. Чистые/зашумленные стерео речевые данные требуют оценки максимального правдоподобия коррекции векторов. В [Hilger et al, 2006] другой алгоритм, называемый уравнивание на основе квантилей (QE), был разработан для компенсирования несоответствия распределения обучающих и тестовых речевых данных на основе квантилей распределения. Параметры компенсационной функции были оценены за счет минимизации квадрата расстояния между текущими квантилями и обучающими квантилями из фильтрационного банка тоновых частот. SPLICE и QE были проанализированы для распознавания зашумленной речи собрания The Wall Street Journal (WSJ) при различных типах и уровнях шума. Значительные улучшения были получены в чистых и комбинированных учебных сценариях.

1.3. Характеристики адаптации диктора

Учебные данные для независимой от диктора системы обычно включают в себя слова большого количества разных дикторов. Вариации акустических характеристик можно рассматривать состоящими из двух компонентов: внедикторский компонент, соответствующий различным фонетическим классам, и междикторский компонент, соответствующий различным голосовым характеристикам некоторых дикторов. С целью различить фонетические классы мы заинтересовались моделированием внедикторских вариаций больше, чем междикторских. Методы нормализации диктора, работающие в области характеристик, нацеленной на подготовку канонической характеристики пространства для устранения, на сколько возможно, изменчивости внедикторских вариаций. [Савченко, Васильев, 2014] Ниже представлены примеры описанных методов:

1) деформации оси частот, чтобы соответствовать длине вокального тракта контрольного громкоговорителя, как в длине тракта вокальной нормализации (VTLN) [Lee et al, 1998];

2) аффинное преобразование функции максимизации вероятности в рамках нынешней модели как пространство характеристик максимального правдоподобия линейной регрессии (fMLLR) [Saon et al, 2000];

3) полноразмерное нелинейное преобразование эмпирического распределения адаптированных данных в соответствии с нормальным распределением ссылки, как Гауссово пространство характеристик.

Далее акустическая модель будет обучаться на каноническом пространстве признаков, которые в идеале становятся лишены междикторской вариации.

1.4. Дискриминативные характеристики

Другим мощным инструментом в арсенале моделирования современных LVCSR систем является в пространственно-характеристическое дискриминативное обучения. Область характеристик минимальных фоновых ошибок (fMPE) [Povey et al, 2005] есть преобразование, обеспечивающее независящие от времени сдвиги регулярных характеристических векторов. Сдвиги получаются путем линейного проецирования из пространства высокой размерности Gaussian posteriors. Проекция обучается таким образом, чтобы повысить уровень распознавания между верными и некорректными последовательностями слов. В сочетании с моделированием области отличительных признаков данный метод обычно приводит к относительному улучшению производительности распознавания на 25% в некоторых задачах. Другой перспективной задачей в извлечении дискриминативной характеристики является использование нейросетевой (NN) параметризации речевого сигнала. Модели, построенные на нейронной сети с акустическими характеристиками, обеспечивают увеличение производительности LVCSR благодаря комбинации систем [Vergyri et al, 2008].

Подводя итог, на рисунке 3 изображен типичный ход предварительной обработки современных систем LVCSR.

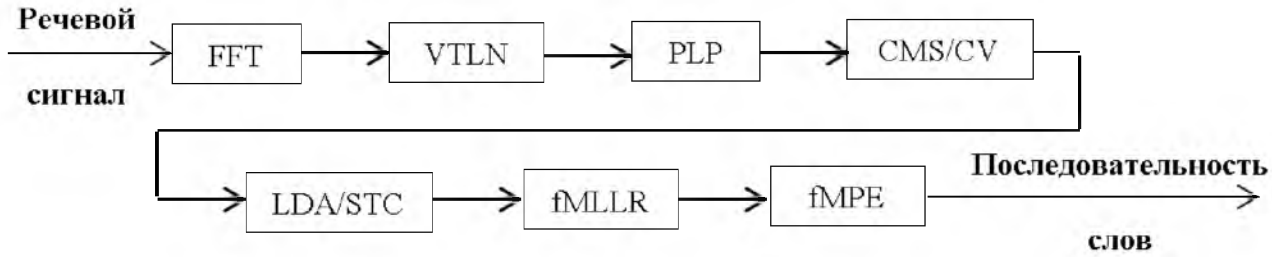

Рис. 3. Этапы предварительной обработки речевого сигнала
Fig.3. Overview of front-end processing steps

2. Акустическое моделирование

2.1. Скрытые модели Маркова

Скрытые модели Маркова (HMMs) [Rabiner, 1989] популярны для представления временных или пространственных последовательностей, например, речи, изображения, видео, текста, музыки, в области биологии и финансов, а также многих других. Предположим, что для акустического моделирования собрано множество *D*-мерных непрерывных многозначных векторов речевых характеристик $X = \{x_t\}_{t=1}^{T}$. Состояние функции плотности вероятности наблюдения вектора признаков x_t в момент времени *t* выражается гауссовой смесью (GMM)

$$p(x_t \mid \Lambda_t) = \sum_{k=1}^{K} \omega_{tk} N(x_t; \mu_{tk}, \sum_{tk}) \quad (1)$$

где параметры состояния $\Lambda = \{\Lambda_t\} = \{\omega_{tk}, \mu_{tk}, \sum_{tk}\}$ включают в себя смесь весов ω_{tk}, средних векторов μ_{tk} и ковариантных матриц $\sum_{tk}$ для *K* компонентов гауссовой смеси. Обычно предполагается, что $\sum_{tk}$ – диагональная, хотя были предложены более сложные модели, такие как подпространство точности и значения (SPAM) [Axelrod et al, 2002], с целью преодоления разрыва между полным и диагональным ковариационным моделированием.

Совместная вероятность коллекции речевых данных *X* задается как

$$p(X \mid \Lambda) = \sum_{S=\{s_1,\ldots,s_T\}} \left[\pi_{s_1} p(x_1 \mid \Lambda_{s_1}) \prod_{t=2}^{T} a_{s_{t-1} s_t} p(x_t \mid \Lambda_{s_t}) \right] \quad (2)$$

Параметры HMMs $\Lambda = \{\pi_t, a_{ij}, \omega_{tk}, \mu_{tk}, \sum_{tk}\}$ подчиняются ограничениям начальных состояний вероятностей $\sum_t \pi_t = 1$, вероятности переходного состояния $\sum_j a_{ij} = 1$ и весам смеси $\sum_K \omega_{tk} = 1$.

2.2. Оценка максимального правдоподобия

Обычные HMMs порождены моделями, обученными по критерию максимального правдоподобия (ML), где параметры модели оцениваются по максимизации совместной функции правдоподобия $p(X \mid \Lambda)$. Оценка ML страдает от проблемы неполноты данных, так как метки состояния $s_t = i$ отсутствуют в целевой функции $p(X \mid \Lambda)$. Алгоритм ожидания максимизации (EM) [Dempster et al, 1977] применяется для решения данной проблемы посредством максимизации функции ожидания или вспомогательной функции логарифмического правдоподобия $\log p(X \mid \Lambda)$ над недостающими переменными $\{S = \{s_t\}\}$. На каждом шаге итерации EM новый критерий ML оценки Λ получается за счет максимизации вспомогательной функции $Q(\Lambda \mid \Lambda^{(k)})$ с учетом предыдущей оценки $\Lambda^{(k)}$ на *k*-ой итерации

$$\Lambda^{(k+1)} = \arg\max_{\Lambda} Q(\Lambda \mid \Lambda^{(k)}) = \arg\max_{\Lambda} \sum_{S} p(S \mid X, \Lambda^{(k)}) \log p(X, S \mid \Lambda) \quad (3)$$

Выполнение EM итераций гарантирует, что функция правдоподобия не будет убывать, т.е. оценки: новая Λ и предыдущая $\Lambda^{(k)}$, - удовлетворяют условию: $p(X \mid \Lambda) \geq p(X \mid \Lambda^{(k)})$, если $Q(\Lambda \mid \Lambda^{(k)}) \geq Q(\Lambda^{(k)} \mid \Lambda^{(k)})$ [Dempster et al, 1977]. На рисунке 4 изображен обзор методик state-of-the-art акустического моделирования для LVCSR.

Рис. 4. Методики акустического моделирования
Fig.4. Techniques of acoustic modeling

В рамках данной статьи будут рассмотрены методы первой группы.

2.3. Дискриминативное обучение

Оценка критерия ML гарантирует «оптимальность» в распределении для порождающей модели. Однако для общих систем распознавания образов желательна "оптимальность" в точности классификации. Будучи непосредственно связанным с точностью классификации, дискриминативное обучение эффективнее, чем оценка критерия ML. В системах LVCSR мы стремимся найти лучшую дискриминативную акустическую модель для достижения наименьшего уровня WER на скрытых тестовых данных. Непосредственная минимизация WER трудна, потому что целевая функция не дифференцируема, что не позволяет применить методы на основе градиента. Альтернативное решение заключается в оценке дискриминативной модели по минимизации частоты ошибок классификации (MCE), являющейся гладкой аппроксимацией коэффициента ошибок слова или предложения. Оценка MCE появилась из правила принятия решений Байеса и значительно опережает оценку ML в распознавании речи [Juang et al, 1997]. Кроме того дискриминативные акустические системы могут быть обучены в соответствии с критерием максимизации взаимной информации (MMI), который выражается как взаимная информация между данными наблюдения X и последовательностью эталонных слов W^r

$$F_{MMI}(\Lambda) \stackrel{\Delta}{=} I_\Lambda(X, W^r) = \log \frac{p_\Lambda(X, W^r)}{p_\Lambda(X) p_\Lambda(W^r)} = \\ = \log p_\Lambda(X \mid W^r) - \log \sum_W p_\Lambda(X \mid W) p(W) \stackrel{\Delta}{=} F^{num}(\Lambda) - F^{den}(\Lambda), \quad (4)$$

или, что эквивалентно, как разница между функцией числителя $F^{num}(\Lambda)$, соответствующей ссылке на последовательность слов W^r, и функцией знаменателя $F^{den}(\Lambda)$ для всех возможных последовательностей слов $\{W\}$. Когда точная ссылка недоступна, декодирующий выход (без контроля со стороны обучения) или соглашение между декодированным входом и какой-то доступной стенограммой могут быть заменены. Знаменатель $F^{den}(\Lambda)$ может быть эффективно аппроксимирован ограниченной суммой только тех последовательностей слов, которые возникают в словесной решетке альтернативных гипотез предложений, полученные путем декодирования слабой (обычно, unigram) языковой модели. Объект в (4) схож с отрицательной ошибкой неправильной классификации функции показателей в оценке MCE. Оценка MMI параметров HMM Λ обычно осуществляется с помощью расширенного алгоритма Баума-Уэлча, максимизирующего «слабый смысл» вспомогательной функции $Q(\Lambda \mid \Lambda^{(k)})$ из [Povey, Woodland, 2002]

$$\sum_S p(S \mid X, W^r, \Lambda^{(k)}) \log p(X, S \mid \Lambda) - \sum_S \sum_W p(S, W \mid X, \Lambda^{(k)}) \times \\ \times \log p(X, S \mid \Lambda) + Q^{sm}(\Lambda \mid \Lambda^{(k)}), \quad (5)$$

где первый и второй члены соответствуют вспомогательным функциям для числителя $F^{num}(\Lambda)$ и знаменателя $F^{den}(\Lambda)$ соответственно, а $Q^{sm}(\Lambda|\Lambda^{(k)})$ обозначает функцию сглаживания, добавляемую для гарантии того, что целевая функция $Q(\Lambda|\Lambda^{(k)})$ увеличится после обновления параметров. Популярная функция сглаживания задается суммой отрицательных расхождений Кульбака-Лейблера между устойчивыми распределениями для Λ и $\Lambda^{(k)}$. Из (4) обучение MMI может интерпретироваться как максимизация log posterior вероятности $\log p_\Lambda(X|W^r)$ правильной последовательности слов W^r [Povey, Woodland, 2002], которая также известна как оценка условий максимального правдоподобия (CML).

В другом подходе дискриминативное обучение, базирующееся на критерии минимума фоновых ошибок (MPE) [Povey, Woodland, 2002]. В отличие от MCE и минимума целевых функций ошибок слова, MPE обучение стремится свести к минимуму взвешенный коэффициент фоновых ошибок или, что эквивалентно, максимизации взвешенной точности фона

$$F_{MPE}(\Lambda)=\sum_{r=1}^{R}\sum_{W}p_\Lambda^k(W|X_r)A(W,W^r), \tag{6}$$

где $X=\{X_r\}_{r=1}^R$ обозначает R учебных предложений, $p_\Lambda^k(W|X_r)$ определяется как масштабируемая posterior вероятность предложения с гипотетической последовательностью слов W скалярной величины k, а $A(W,W^r)$ – число корректных фонов в W (взятых из эталонной словесной последовательности W^r). MPE обучение приводит к улучшению точности по сравнению с ML и MMI обучением в различных задачах LVCSR [Povey, Woodland, 2002]. MPE обучение может быть просчитано в решетке основ, где решетка или словесная диаграмма генерируется для эффективного кодирования всех возможных последовательностей слов, которые имеют поддающиеся оценке вероятности, полученные из акустических данных. Вид MPE, называемый минимальной ошибкой фонового кадра (MPFE), был предложен в [Zheng , Stolcke, 2005] и имеет то преимущество, что он используется для точности измерения фоновых кадров, которую легче вычислить.

В дополнении к пространственной модели дискриминативного обучения для Λ параметров HMM похожая целевая функция, либо MPE, либо MMI, может быть оптимизирована для выполнения дискриминативного обучения на пространстве признаков, состоящего из оценки матрицы проекции, которая отображает заданные векторы высокой размерности в смещенные вектора, которые добавляются к акустическим признакам [Povey et al, 2005]. Более конкретно обучение в пространстве признаков MPE (fMPE) или пространстве признаков MMI (fMMI) выполняется путем преобразования акустических свойств x_t в $\hat{x}_t=\{\hat{x}_{td}\}$ для каждого кадра t следующим образом $\hat{x}_t=x_t+Mh_t$, где $M=\{m_{dj}\}$ – матрица трансформации, а $h_t=\{h_{tj}\}$ – вектор признаков высокой размерности, который сформирован из Gaussian posteriors, взятого для отдельного кадра и вычисленный из GMM. Матрица трансформации M оценивается по максимизации вспомогательной функции $Q(\Lambda|\Lambda^{(k)})$ (без сглаживания членов) по тому же критерию, что использовались в (4) и (6), применяя алгоритм градиентного спуска

$$m_{dj}\leftarrow m_{dj}+v_{dj}\frac{\partial Q}{\partial m_{dj}}=m_{dj}+v_{dj}\sum_t\frac{\partial Q}{\partial\hat{x}_{td}}h_{tj}, \tag{7}$$

где параметр конкретной скорости обучения v_{dj} определяется эмпирически. Тогда как целевые функции MPE и MMI зависят от Λ параметров HMM и трансформированных признаков $\{\hat{x}_t\}$, частная производная в (7) содержит прямую и смешанную производные

$$\frac{\partial}{\partial\hat{x}_t}Q(\hat{x}_t,\lambda(\hat{x}_t)|\Lambda^{(k)})=\underbrace{\frac{\partial Q}{\partial\hat{x}_t}}_{\text{прямая}}+\underbrace{\frac{\partial Q\partial\Lambda}{\partial\Lambda\partial\hat{x}_t}}_{\text{смешанная}}, \tag{8}$$

подробно описанные в [Povey et al, 2005]. Заметим, что fMPE может быть записано эквивалентно как смесь зависимых от времени смещений $\hat{x}_t=\sum_j h_{tj}(x_t+m_j)$, где h_{tj} – posterior Гауссиана j в момент времени t, m_j – j-ый столбец матрицы M. Обобщение fMPE, называемое ограниченным по области линейным преобразованием (RDLT), приводится в [Zhang et al, 2006]. Оно включает в себя замену

смещений на смесь аффинных преобразований $\hat{x}_t = \sum_j h_{jt}\left(A_j x_t + b_j\right)$. В применении к некоторым задачам LVCSR fMPE обучение превосходит MPE обучение. Производительность системы дополнительно улучшена путем объединения fMPE и MPE обучения параметров модели (обозначается как fMPE + MPE) [Povey et al, 2005].

В еще одном подходе, основанном на методах с большим запасом классификации, увеличенная целевая функция в MMI (BMMI) получается за счет введения параметра масштабирования K и увеличения коэффициента внутри целевой функции MMI из (4) как показано ниже [Saon , Povey, 2008]

$$F_{BMMI}(\Lambda) \stackrel{\Delta}{=} \sum_{r=1}^{R} \log \frac{p_\Lambda^K\left(X_r \mid W^r\right) p\left(W^r\right)}{\sum_W p_\Lambda^K\left(X_r \mid W\right) p(W) \exp\left(-bA\left(W, W^r\right)\right)}. \quad (9)$$

Увеличение коэффициента контролируется параметром b и точностью фоновой оценкой $A\left(W, W^r\right)$ между предполагаемой и эталонной последовательностями слов $\left(W, W^r\right)$. Вышеизложенная идея BMMI обучения применяется для искусственного увеличения вероятности наиболее «запутанных» предложений, которые имеют больше ошибок, так что алгоритм обучения уделяет им больше внимания. Пространственно-характеристическое и пространственно-модельное BMMI обучение (записывается как fBMMI+BMMI), превосходит fMPE+MPE в решении нескольких задач LVCSR [Zheng, Stolcke, 2005, Saon , Povey, 2008], и в настоящее время является лучшей дискриминативной обучающей схемой для LVCSR из известным нам.

Для того чтобы сделать связь с большим запасом классификации более явной, в [Saon, Povey, 2008] и [Saon et al, 2009] был модифицирован критерий BMMI как

$$F_{PLM}(\Lambda, b) \stackrel{\Delta}{=} b + \frac{1}{\rho} \sum_{r=1}^{R} \log \frac{p_\Lambda\left(X_r \mid W^r\right) p\left(W^r\right)}{\sum_W p_\Lambda\left(X_r \mid W\right) p(W) \exp\left(bH\left(W, W^r\right)\right)}, \quad (10)$$

что рассматривается как критерий penalized с большим запасом классификации (PLM) и берет свое начало из задачи ограничительной оптимизации для общей классификации с большим запасом

$$\max b$$
$$s.t. \_ \log p_\Lambda\left(X_r \mid W^r\right) - \log\left(X_r, W\right) \geq bH\left(W, W^r\right), \quad \forall W, 1 \leq r \leq R. \quad (11)$$

В (10) и (11) $H\left(W, W^r\right)$ обозначает число кадров фоновых ошибок или промежуток Хемминга между W и W^r, $b \geq 0$ рассматривается в качестве параметра предельного (маржинального) масштабирования, ρ - penalty параметр, контролирующий компромисс между предельной максимизацией и ограничениями. Контролирующий параметр аналогичен тому, что принят в мягкой предельной классификации [Bishop, 2006], где мягкая маржинальность пропорциональна числу ошибок в предполагаемом предложении. В [Bishop, 2006] оценка большой предельности была предложена для выполнения выбора кадра и выборки высказывания. Поддерживаемые маркеры для акустического моделирования определены аналогично опорным векторам, используемым в машинах с опорными векторами. В [Chen, Chien, 2009] оценка большой маржинальности Бейса была предложена для комбинирования обучения Бейса и оценки большой маржинальности HMM обучения, а также модели регуляризации. Дискриминативное обучение на пространстве признаков и обучение большой предельности, базирующееся на fMPE, fMPE+MPE, BMMI, fBMMI+BMMI и PLM, оказались эффективными в повышении производительности LVCSR.

Работа выполнена при поддержке базовой части государственного задания СКФУ №2563.

Список литературы
References

Савченко В.В., Васильев Р.А. 2014. Анализ эмоционального состояния диктора по голосу на основе фонетического детектора лжи. Научные ведомости БелГУ. Сер. История. Политология. Экономика. Информатика. 21(192): 186-195.

Savchenko V.V., Vasiliev R.A. 2014. Analiz jemocional'nogo sostojanija diktora po golosu na osnove foneticheskogo detektora lzhi. Nauchnye vedomosti BelGU. Ser. Istorija. Politologija. Jekonomika. Informatika [The analysis of the emotional condition of the announcer on the voice on the basis of the phonetic lie detector] 21(192): 186-

195.

Axelrod S., Gopinath R., Olsen P. 2002. Modeling with a subspace constraint on inverse covariance matrices. Proc. Int. Conf. Spoken Language Processing (ICSLP): 2177–2180.

Bishop C. M. 2006. Pattern Recognition and Machine Learning. New York: Springer-Verlag.

Chen J.-C., Chien J.-T. 2009. Bayesian large margin hidden Markov models for speech recognition. Proc. Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP): 3765–3768.

Chen S., Kingsbury B., Mangu L., Povey D., Saon G., Soltau H., Zweig G. 2006. Advances in speech transcription at IBM under the DARPA EARS program. IEEE Trans. Speech Audio Processing. 14 (5): 1596–1608.

Dempster A. P., Laird N. M., Rubin D. B. 1977. Maximum likelihood from incomplete data via the EM algorithm. J. R. Stat. Soc. B. 39 (1): 1–38.

Deng L., Acero A., Plumpe M., Huang X. 2000. Large-vocabulary speech recognition under adverse acoustic environments. Proc. Int. Conf. Spoken Language Processing (ICSLP): 806–809.

Furui S. 1986. Speaker independent isolated word recognition using dynamic features of speech spectrum. IEEE Trans. Acoust., Speech, Signal Processing. 34 (1): 52–59.

Gales M. J. F. 1998. Maximum likelihood linear transformations for HMM-based speech recognition. Comput. Speech Lang. 12 (2): 75–98.

Hermansky H. 1990. Perceptual linear predictive (PLP) analysis of speech, J.Acoust. Soc. Am. 87 (4): 1738–1752,.

Hilger F., Ney H. 2006. Quantile based histogram equalization for noise robust large vocabulary speech recognition. IEEE Trans. Audio Speech Lang. Processing. 14 (3): 845–854.

Juang B.-H., Chou W., Lee C.-H. 1997. Minimum classification error methods for speech recognition. IEEE Trans. Speech Audio Processing. 5 (3): 257–265.

Lee L., Rose R. 1998. A frequency warping approach to speaker normalization. IEEE Trans. Speech Audio Processing. 6 (1): 49–60.

Povey D., Kingsbury B.,. Mangu L, Saon G., Soltau H., Zweig G. 2005. fMPE: Discriminatively trained features for speech recognition. Proc. Int. Conf. Acoustic, Speech, and Signal Processing (ICASSP): 961–964.

Povey D., Woodland P. C. 2002. Minimum phone error and I-smoothing for improved discriminative training," in Proc. Int. Conf. Acoustic, Speech, and Signal Processing (ICASSP): 105–108.

Rabiner L. R. A tutorial on hidden Markov models and selected applications in speech recognition. Proc. IEEE, vol. 77, no. 2, pp. 257–286, 1989.

Saon G., Padmanabhan M., Gopinath R., Chen S.. Maximum likelihood discriminant feature spaces. Proc. Int. Conf. Acoustic, Speech, and Signal Processing (ICASSP), 2000, pp. 1129–1132.

Saon G., Povey D. 2008. Penalty function maximization for large margin HMM training. Proc. Annu. Conf. Int. Speech Communication Association (INTERSPEECH): 920–923.

Saon G., Povey D., Soltau H. 2009. Large margin semi-tied covariance transforms for discriminative training. Proc. Int. Conf. Acoustic, Speech, and Signal Processing (ICASSP): 3753–3756.

Vergyri D., Mandal A., Wang W., Stolcke A., Zheng J., Graciarena M., Rybach D., Gollan C., Schlueter R., Kirchhoff K., Faria A., Morgan N. 2008. Development of the SRI/Nightingale Arabic ASR system. Proc. Annu. Conf. Int. Speech Communication Association (INTERSPEECH): 1437–1440.

Zhang B., Matsoukas S., Schwartz R. 2006. Discriminatively trained region dependent feature transforms for speech recognition. Proc. Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP): 313–316.

Zheng J., Stolcke A. 2005. Improved discriminative training using phone lattices. Proc. Annu. Conf. Int. Speech Communication Association (INTERSPEECH): 2125–2128.