Вычислительный центр им. А. А. Дородницына
Российской Академии Наук

На правах рукописи

Нгуен Минь Туан

РАЗРАБОТКА АЛГОРИТМОВ ПОСТРОЕНИЯ ОЦЕНОК ДОСТОВЕРНОСТИ ДЛЯ СИСТЕМ РАСПОЗНАВАНИЯ РЕЧИ

Специальность 05.13.11 - математическое и программное
обеспечение вычислительных машин,
комплексов и компьютерных сетей

Диссертация на соискание ученой степени
кандидата технических наук

Научный руководитель в.н.с, к.ф.-м.н. Чучупал
Владимир Яковлевич

Москва, 2008

Содержание

Введение

Прогресс современного общества в значительной мере обусловлен развитием автоматических и роботизированных систем. Компьютеры и микропроцессоры стали неотъемлемым атрибутом жизни людей в индустриально развитых странах. Научно-техническая проблема создания адекватных средств для взаимодействия человека с компьютерными системами приобрела в последние десятилетия важный социальный статус.

Одним из наиболее очевидных и перспективных путей организации взаимодействия человека с компьютером является использование человеческой речи, в частности, автоматическое распознавание речевых сообщений. Исследования по автоматическому распознаванию речи начались более пятидесяти лет назад, в середине прошлого века [19] и интенсивно продолжаются в настоящее время.

Первоначально основной целью автоматического распознавания речи была разработка методов точного преобразовании акустического речевого сигнала в текстовое сообщение для создания так называемой «фонетической пишущей машинки» [8].

С течением времени, с учетом опыта практической реализации систем распознавания речи, произошла переоценка целей и задач этой научной области, на передний план вышли вопросы распознавания и понимания естественной речи, а также создания диалоговых систем. В таких условиях наблюдаемый речевой сигнал может содержать, помимо известных системе слов, также различные акустические события, например, незнакомые слова, обрывки речи, кашель, смех и т.п.

Сейчас задача автоматического распознавания речи трактуется как преобразование речевых сообщений в адекватную речевому высказыванию последовательность действий, в том числе, орфографическую запись высказывания. Для диалоговых систем, например, систем резервирования

билетов на транспорт, систем управления бортовой аппаратурой самолета или робототехнического устройства точная текстовая запись высказывания, вообще говоря, не требуется, здесь важно понять значения отдельных терминов. Например, для систем резервирования авиабилетов это могут быть имена пунктов вылета и прилета, дата и время полета.

Успехи в создании методов и технологий распознавания речи очевидны. С точки зрения известного японского специалиста С. Фуруи [28] наиболее значимыми научными и технологическими результатами, полученными за последние годы являются:

- переход от распознавания на основе шаблонов слов к статистическому моделированию речи с помощью Скрытых Марковских Моделей и п-грамм.
- переход от мер сходства на основе расстояний к мерам близости на основе правдоподобия
- использование дискриминантных методов для распознавания речи
- использование контекстно-зависимых акустических моделей звуков
- переход от распознавания изолированно произносимых слов к распознаванию слитной речи
- переход от систем распознавания с небольшими словарями к системам со словарями в десятки тысяч слов
- распознавание речи в условиях телефонного канала
- распознавание речи произвольного человека
- распознавание естественной речи
- распознавание речи в ситуациях полилогов
- понимание речевых сообщений
- развитие мультимодальных систем распознания речи
- реализация сложных систем распознавания целиком на уровне программного кода
- развитие специального программного обеспечения, его стандартизация

- появление коммерчески успешных продуктов с использованием распознавания речи Успехи, достигнутые научными коллективами, можно количественно измерить результатами (например, в терминах основной характеристики эффективности систем распознавания речи - вероятности пословной ошибки распознавания), которые получены при решении специально выбранных тестовых заданий. В следующей таблице приведены вероятности пословной ошибки распознавания для лучших лабораторных систем распознавания речи, которые были получены при испытаниях на четырех индикативных проблемно-ориентированных задачах. Для сведения также приведены характеристики сложности задач - размер словаря и перплексия (коэффициент ветвления) языка [40].

Таблица 1

Характеристики нескольких современных систем распознавания речи

Задача	Размер словаря	Перплексия языка	Вероятность ошибки
Распознавание слитно произносимых цифр	11	11	0.5%
Деловые новости (читаемая речь)	20000	200	3%
Новости (читаемая речь)	64000	-	10%
Телефонные разговоры	64000	-	20%

Из представленных данных следует, что распознавание естественной произвольной речи, тем более в ограниченном по полосе частот, канале передачи, каким является телефонный канал, далеко от удовлетворительного: каждое пятое слово распознается неправильно. В этом нет ничего необычного, поскольку распознавание речи у человека неразрывно связано с

ее пониманием и мультимодальной обработкой, то есть анализом смысла высказывания, учетом контекстной информации, мимики и т.п.

Основная причина относительно невысокой эффективности систем речевой технологии заключается в вариативности речевого сигнала, которая обуславливается, например, индивидуальными особенностями дикторов, характеристиками каналов связи, а также влиянием окружающей обстановки.

На эффективность автоматического распознавания речи также оказывают существенное влияние условия прикладной области, в частности, размер словаря. Как правило, словарь системы распознавания является замкнутым, то есть содержит все слова, которые могут быть произнесены и должны быть распознаны. Увеличение размера словаря, вообще говоря, снижает вероятность правильного распознавания.

Потребность распознавания естественной, неограниченной, по словарному составу, речи, приводит к тому, что требование правильного распознавания всего высказывания вряд ли осуществимо и обычно не требуется. Поскольку в данном случае словарь системы является открытым, необходимо предусмотреть возможность отказа системы от распознавания каких-то частей речевого высказывания, которые содержат новые, не входящие в словарь системы, выражения и слова. Таким образом, появляется необходимость решения проблемы идентификации в речевом потоке новых, так называемых, несловарных (OOV, «out of vocabulary») слов или иных акустических событий. Естественным способом решения этой проблемы является синтез так называемых оценок достоверности для результатов распознавания, на основе значений которых можно, в частности, идентифицировать OOV.

Под оценкой достоверности (английский термин «confidence measure») для некоторого результата распознавания речи, под которым может подразумеваться отдельное слово, звук или предложение, здесь и далее будет

пониматься число, в интервале от 0 до 1, которое характеризует степень доверия или уверенности в правильности этого результата.

Применение оценок достоверности также может повысить эффективность использования традиционных систем распознавания речи, оперирующих с замкнутыми словарями. Часто эти системы используются как составная часть более крупных автоматических систем, например, управления робототехническими комплексами, доступа к информационным ресурсам, диалоговых систем. В этом случае существует возможность коррекции ошибок автоматического распознавания речи на основе дополнительной информации, которой располагает система верхнего уровня. Такая коррекция будет более успешна, если система распознавания речи предоставит расширенную информацию о результате распознавания, включающую не только предполагаемые слова, но и оценку их достоверности.

Важность решения проблемы построения эффективных оценок достоверности для систем распознавания речи увеличивается по мере дальнейшего прогресса в области речевых технологий. Это обстоятельство определяет **актуальность исследований** в этом направлении.

Цель диссертационной работы заключается в исследовании и разработке эффективных алгоритмов построения оценок достоверности для систем автоматического распознавания речи.

Достижение указанной цели предполагает решение следующих **основных задач:**

1. Исследование существующих методов моделирования и автоматического распознавания речи, а также известных методов построения оценок достоверности для систем распознавания речи.
2. Разработка новых методов и алгоритмов построения оценок достоверности результатов работы систем распознавания речи.

3. Программная реализация предлагаемых алгоритмов и проведение экспериментальных исследований их эффективности.

В качестве **методов исследования** использовались методы математического анализа, методы цифровой обработки сигналов, теории распознавания образов, теории вероятностей, методы кластеризации, теории оптимизации, теории формальных языков.

Научная новизна заключается в том, что предложен новый метод построения оценок достоверности для систем распознавания речи, который основан на построении дополнительных моделей распределения признаков речевого сигнала. Разработаны алгоритмы оценивания значений параметров дополнительных моделей, а также выбора оптимального количества их параметров.

Практическая ценность диссертации. Предложенный метод формирования оценок достоверности показал высокую эффективность при верификации результатов распознавания речи. Исследования были выполнены в рамках работ по проектам «Разработка и тестирование системы распознавания речевых команд управления в акустико-фоновой обстановке кабины пилота» и «Разработка и исследование методов распознавания речи на основе комбинированных моделей звуков» (гранты РФФИ № 06-08-1534 и №07-01-00657).

Основные научные результаты диссертации, выносимые на защиту:

1. Метод формирования оценок достоверности для систем распознавания речи, основная идея которого заключается в построении специальных (дополнительных) моделей распределения векторов признаков речевого сигнала.
2. Алгоритм оценивания параметров дополнительных моделей распределения по обучающей выборке
3. Алгоритм выбора оптимального количества параметров дополнительных моделей.

Апробация работы. Результаты диссертация докладывались на XII международной конференции «Речь и Компьютер» SPECOM'2007 (Москва, 2007 г.), на XIX сессии Российского Акустического Общества (Нижний Новгород, 2007 г.), на XIII всероссийской конференции «Математические методы распознавания образов» (Санкт-Петербург, 2007 г.), на VII Открытом немецко-российском семинаре «Распознавание образов и понимание изображений» (Эттлинген, 2007 г.), а также на семинаре отдела математических проблем распознавания образов и методов комбинаторного анализа ВЦ РАН (Москва, 2008 г.).

Публикации. По результатам диссертационной работы опубликовано 6 статей в научных изданиях [1-6].

Диссертационная работа состоит из введения, трех глав, заключения, двух приложений и библиографического списка использованных источников. Общий объем составляет 102 страницы, в том числе 13 рисунков и 20 таблиц. Библиографический список включает 85 наименований.

Первая глава диссертации является обзорной. В первом разделе рассмотрен вероятностный подход к моделированию и распознаванию речи. Выделены основные компоненты (модули) современных систем распознавания речи: модуль выделения акустических признаков, акустическая модель, языковая модель и модуль декодирования. Сформулированы основные требования к модулю выделения акустических признаков и описан метод формирования акустических векторов на основе мел-кепстрального анализа, который использован при проведении численных экспериментов в данной работе. Дано определение скрытой Марковской модели, которая используется для построения моделей звуков. Описаны методы построения моделей языка с помощью формальной грамматики и статистических N-грамм. Во втором разделе проведен анализ существующих методов формирования и измерения эффективности оценок достоверности результатов автоматического распознавания речи. Оценки достоверности

условно разделены на 3 группы на основе способов их формирования. Первая группа состоит из т. н. элементарных характеристик, которые получаются в процессе распознавания речи. В качестве общеупотребительных примеров таких характеристик рассмотрены акустические оценки и плотность гипотез. Оценки достоверности второй группы основываются на апостериорной вероятности наблюдения распознанного слова при заданном наборе акустических векторов. Описаны методы оценки достоверности на основе вычисления априорной вероятности с помощью графа слов (Word Graph). К третьей группе относятся оценки достоверности на основе отношения правдоподобия с использованием специальных акустических моделей. Приведены описания оценок достоверности третьей группы. Для каждой из рассмотренных групп оценок приведены их положительные и отрицательные стороны, приведены численные результаты экспериментов.

Вторая глава посвящена описанию предлагаемых автором методов и алгоритмов формирования оценок достоверности. В первом разделе определены модели распределений векторов признаков, названные целевыми и альтернативными моделями. Описан метод формирования оценок достоверности для гипотезированных слов, который основан на использовании отношений правдоподобия для целевых и альтернативных моделей. Сформулирован критерий обучения целевых и альтернативных моделей. Показано, что обучение этих моделей можно проводить методом градиентного спуска и указаны недостатки такого подхода. Описан новый алгоритм обучения моделей, который свободен от указанных недостатков.

В третьей главе приведены результаты практического применения предложенных в работе методов и алгоритмов. В первом разделе дано описание корпуса речевых данных FaVoR. На основе данных корпуса FaVor сконструированы три выборки данных: обучающая, настроечная и тестовая. Обучающая выборка, предназначена для обучения моделей звуков. Настроечная выборка используется для обучения целевых и альтернативных

моделей. На тестовой выборке производилась оценка эффективности предложенных методов формирования оценок достоверности. Приведены характеристики каждой из выборок данных. Во втором разделе описаны модули базовой системы распознавания речи, основанной на вероятностном подходе. Модуль извлечения векторов признаков преобразует входной речевой сигнал в последовательность векторов признаков, состоящих из мел-кепстральных коэффициентов, их первых и вторых производных. Для акустического моделирования речевого потока был выбран подход на основе построения т.н. контекстно-зависимых моделей звуков речи, которые моделировались с помощью СММ. Приведены результаты работы базовой системы распознавания на настроечной и тестовой выборок. В третьем разделе приведены численные результаты применения предложенных в работе методов и алгоритмов. Сравнена эффективность предложенного метода формирования оценок достоверности с другими методами, для которых опубликованы численные значения оценок эффективности . Показано, что предложенные в диссертации алгоритмы позволяют существенно снизить вероятности пропуска правильно распознанных слов и вставки неправильно распознанных слов при работе системы распознавания речи.

В Заключении сформулированы основные результаты, полученные в ходе работы над диссертацией.

Глава 1. Обзор современных методов распознавания речи и оценивания достоверности результатов распознавания

1.1 Вероятностный подход к моделированию и распознаванию речи

В настоящее время вероятностный подход является общепринятым при моделировании и распознавании речи. Пусть дана последовательность векторов признаков *X,* которая соответствует речевому высказыванию. Задача автоматического распознавания речи может быть сформулирована так: найти самое вероятное предложение (цепочку слов) *S** языка *L,* которое соответствуете [11], т.е.

$$S^* = \text{argmax} P(S\backslash X) \qquad \text{n}_{\text{n}}$$

Согласно формуле Байеса вероятность *P(S\X)* можно представить в следующем виде

$$P(X) \qquad {}^{\kappa} \quad \}$$

Подставляя (1.2) в (1.1), имеем

$$^{S} = \underset{Set}{\text{argmax}} \quad \underset{\Gamma(\Pi)}{\underbrace{P(X\backslash S)P(S)}} \qquad (1.3)$$

Так как поиск осуществляется среди всех возможных предложений языка *L,* то вероятность *P(X)* является инвариантным значением относительно выбора предложения *S.* Таким образом, формула (1.3) сводится к виду

$$S^* = \underset{Set}{\text{argmax}}\ \text{P(X} \mid S)P(S) \qquad {}^{(14)}_{\text{V}}\ ''$$

Два сомножителя в (1.4) соответствуют двум основным компонентам, или модулям систем распознавания речи: акустико-фонетического моделирования, который предназначен для оценки вероятности $P(X \setminus S)$ и моделирования языка, который предназначен для оценки вероятности $P(S)$ каждого гипотетического предложения S. Также в систему распознавания речи входят модуль параметризации, который служит для извлечения векторов признаков X из входного речевого сигнала Y, и декодирования, который осуществляет поиск оптимального (самого вероятного) предложения среди всевозможных предложений языка L. Типичная структура системы распознавания речи представлена на рис. 1. В следующих разделах описаны функции и алгоритм работы каждого из модулей.

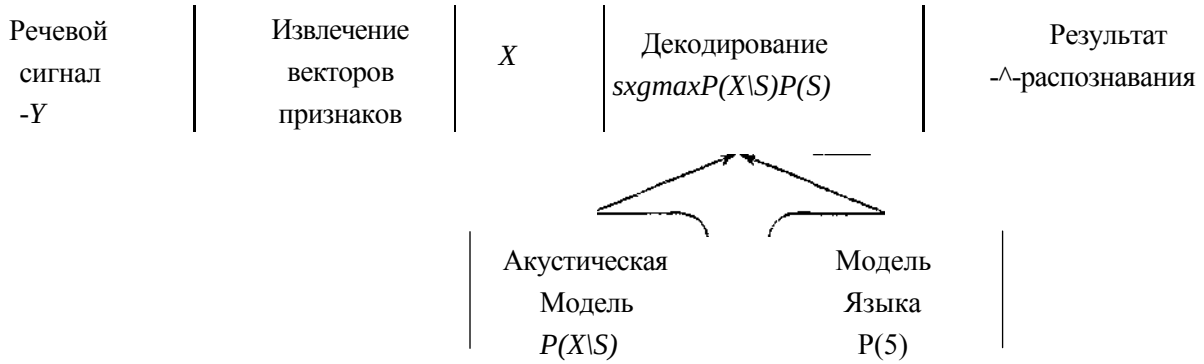


Рисунок 1. Принципиальная структура системы автоматического распознавания речи

1.1.1 Извлечение признаков речевого сигнала

Извлечение признаков речевого сигнала является первым этапом при распознавании речи. На этом этапе входной дискретизированный речевой сигнал, представляющий собой последовательность $Y(t)$ длины T ($/ = 1,...,T$), преобразуется в набор векторов признаков X, пригодных для дальнейшего анализа и обработки. Признаки выбираются таким образом, чтобы разные фонемы имели различимые значения признаков. В то же время,

желательно, чтобы вариация входного сигнала одного и той же фонемы не влияла существенным образом на значения её признаков.

Существует довольно много различных систем признаков, например, коэффициенты линейного предсказания речи (LPC, linear prediction coefficients) [10, 35, 47], мел-спектральные (MFFB, mel-frequency filter bank) и мел-кепстральные коэффициенты (MFCC, mel-frequency cepstrum coefficients) [29, 54, 66], вейвлетные (wavelet) [24, 30, 42] и др.

Все экспериментальные результаты и программное обеспечение данной работы получены с использованием мел-кепстральных коэффициентов и их производных по времени. Выбор этих параметров был обусловлен следующими обстоятельствами:

1. по сравнению с параметрами модели линейного предсказания, мел-кепстральные коэффициенты представляются более стойкими к помехам и искажениям входного сигнала. Они не требуют оценки основного тона и могут быть использованы, в том числе, для анализа и распознавания неречевых звуков.
2. по сравнению с мел-спектральными коэффициентами мел-кепстральные менее чувствительны к изменениям амплитудно-частотных характеристик тракта связи как, например, подъем или спад.
3. мел-кепстральные коэффициенты наиболее часто применяются при построении современных систем распознавания речи.

На рис. 2 показана блок-схема алгоритма формирования векторов признаков.

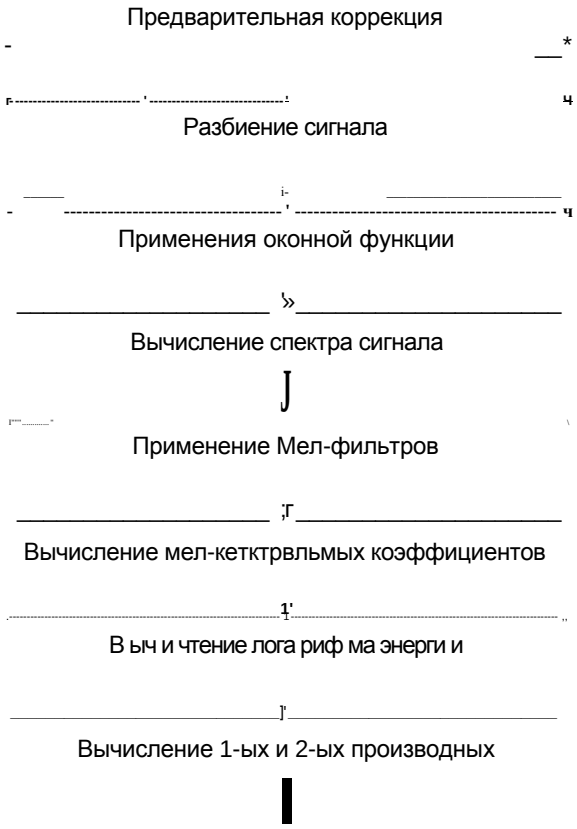


Рисунок 2. Блок-схема алгоритма формирования векторов признаков

Блоки алгоритма выполняют следующие операции:

- Предварительная коррекция. Речевой сигнал *Y(t)* пропускается через фильтр высоких частот

$$Y_2\{t) = Y\{t)-aY\{t-\) ,$$

где *a* - коэффициент коррекции, $0.9 < a < 1$. Этот шаг вызван необходимостью спектрального сглаживания сигнала, который становится менее восприимчивым к различным шумам, возникающим в процессе обработки.

- Выделение кадров анализа сигнала. Сигнал $Y_2(t)$ разбивается на последовательность кадров (сегментов) с равными длинами, и с перекрытием от 1/3 до 1/2 своей длины. Перекрытие используется для предотвращения

потери информации о сигнале на границе. Обычно выбирается длина кадра, соответствующая временному интервалу в 20-30 мс, т.к. на данном интервале речевой сигнал считается стационарным. В результате разбиения получается *К* кадров $Y_2\backslash n),...,Y_2^K(n),0<n<N-l.$

- Обработка кадров. Для подавления нежелательных граничных эффектов, возникающих в результате разбиения, каждый кадр $Y_2^k(n)$ умножается на оконную функцию *w(n)*

$Y_3^k(n) = Y_2^k(n)xw(n)$ В качестве оконной функции *w{ri)* часто используется окно Хэмминга

$$w(\text{»} = 0.54\text{-}0.46\cos \frac{27\text{ГИ}}{\text{---------}} \quad ,0<n<N-l$$

- Следующим шагом является вычисление спектра сигнала в каждом кадре с помощью дискретного преобразования Фурье

$$\text{У/С?}) = \sum_{\text{л=0}}^{N-l} 2\text{X}(\text{»})e^{\frac{2nsn}{N}}, 0 < s < N - l$$

- Оценка амплитудного спектра и моделирование гребенки фильтров с центральными частотами, равноразнесенными по шкале Мелов (Mel Filter Bank). Для этого амплитуды суммируются в частотных полосах, выбранных по шкале Мелов, с весовыми коэффициентами *H(s,m)* и полученная суммарная амплитуда логарифмируется

$$7/\text{»} = \log \sum_{.5=0}^{N-\backslash} \backslash YiY_4^k(s)\backslash H(s,m) \quad \Big| \; Л < m < M$$

где *М* — количество частотных полос.

Мел-фильтры *H{s,m)* задаются формулой

$$H(s,m) = \begin{cases} 0, f_s < f_c(m-\backslash) \\ /,0)-/0-1) \quad ,/_c O-i)</</O) \\ /,-/> + !) \quad ,/O)^/</> + !) \\ /0)-/0+1) \\ [O,/, */> + !) \end{cases}$$

где $/_s$ - частота спектра 7^(^), $f_c(m)$ - частоты, расположенные равномерно по шкале Мелов

$$Me/(/_c(/и)) = 11251п^{\Gamma} \quad 1 + \frac{/O)}{700} л$$

- Дискретное косинусное преобразование (ДКП). Мел-кепстральные коэффициенты получаются в результатепременения дискретного косинусного преобразования к выходамгребенки мел-фильтров (мел-спектру)

$$2XO)\cos_{m=\backslash}^{м} C*(7) = \quad m \text{ ----} \quad ,\backslash<1<M$$

(Ш(
iY\
yMy

- Оценка логарифма энергии сигнала. Кроме кепстральных коэффициентов, в качестве дополнительного элемента вектора признаков используется значение логарифма энергии сегмента

$$£* = \log \sum_{\backslash n=0}^{N-1} 2(1?(»))$$

- Первые и вторые производные коэффициентов. К описанным признакам (мел-кепстральные коэффициенты и логарифм энергии) присоединяются их первые и вторые производные по времени, которые вычисляются по формулам

$$AC^{*}(/) = 2C^{k+2}(/) + C^{k+l}(/) - C^{k}\text{-}^{X}(/) - 2C^{k}\text{-}^{2}(/)\ AAC^{A}(/) = -2C^{i+2}$$
$$(/) + C^{uu}(/) + 2C^{*}(/) + C^{k}\sim^{l}(/) - 2C^{A}\sim^{2}(/)$$
$$A£^{*} = 2£^{i+2} + E^{ы} - E^{k}\sim^{l} - 2E^{l}\sim^{2} AAE^{k} =$$
$$\sim 2E^{k+2} + E^{k+]} + 2E^{k} + E^{k}\sim^{l} - 2E^{1,2}$$

1.1.2 Моделирование речевого сигнала на акустическом уровне

Целью акустической модели является оценка вероятности $P(X \setminus S)$

сигнала X при заданной цепочке слов S. Для системы распознавания речи с большом словарем построение модели для каждой цепочки слов представляется невозможным, так как число допустимых цепочек слов в этом случае огромно. Вместо этого, строятся акустические модели для более мелких речевых единиц, т. н. фонов. Акустическая модель слова получается путем соединения моделей входящих в него фонов. Аналогично, акустическая модель цепочки слов представляет собой конкатенцию акустических моделей слов. В ряде работ [например, 82] было показано, что использование в качестве фонов контекстно-зависимых моделей фонем: бифонов (biphone), и Трифонов (triphone) существенно улучшает характеристики системы распознавания речи.

Существуют несколько подходов к построению акустической модели, например, нейронные сети [27, 31, 61], скрытые Марковские модели (СММ) [43, 56, 58], байесовские сети [21, 85, 70]. Использование СММ является на сегодняшний день наиболее широко применяемым и эффективным подходом к проблеме построения акустической модели.

Скрытая Марковская модель $Я = (A, B,$ П) определяется следующими параметрами

- множество состояний модели 5 = (^,...,^), где N - количество

состояний. Состояние модели в момент времени t обозначается q_t.

- множество различных символов наблюдения $O = \{o_к\}$

- матрица вероятностей переходов между состояниями $A = \{a_{(}\}$

$$Щ] = {}^{P}(Ч< = Sj \text{ I } ?м = \wedge,), 1 < i, j < N .$$

- множество распределений вероятностей появления символов

  наблюдения в состоянии s , $B = \{bj(o)\}$, где

$$Ъ.(o) = P(o \mid q, = Sj), l < j<N, oeO$$

- начальное распределение вероятностей состояний П = {я-.}

$$7t_J = P(q_l = s_J), l£jZN$$

В зависимости от множества символов наблюдения существуют 2 класса СММ: дискретная СММ и непрерывная СММ. Для дискретной СММ множество символов наблюдения представляет собой конечный алфавит из M символов и функции распределения являются дискретными. В случае непрерывной СММ имеется дело с непрерывным множеством символов наблюдения, а функции распределения, как правило, описываются с помощью смесей нормальных распределений

$$Ь\wedge o) = \wedge c_{1к} Щo, \wedge_к, 1Ь;_к \backslash\backslash <] < M ,$$

где M - количество нормальных распределений, описывающих

распределения j-го состояния, c_{jk} - вес k-ой смеси j-го состояния ($c_{Jk} > 0$ и

$\wedge c_{jk} = 1$), $N\{o,/Li_{jk}, Y_к)$ - нормальное распределение с вектором средним $ju_к$

и ковариационной матрицей £ A

$$\wedge^{o'}\wedge^{*'2}\wedge^{=}{}_{\epsilon} \qquad \frac{1}{|1/2^{\epsilon X}P} \qquad (\quad 1 \qquad {}_{m} \quad . \qquad \wedge$$

$$—iP\text{-}/i_{jk}) \qquad I._{jk}(p\text{-}M_{jk})$$

d - размерность символа наблюдения o .

Матрица вероятностей переходов определяет топологию СММ. СММ называется эргодической если все элементы матрицы отличаются от нуля. В задачах распознавания речи часто используется лево-правая модель (модель Бэкиса), где переходы из одного состояния возможны только в то же состояние или последующее состояние. Модель Бэкиса позволяет учитывать временные характеристики речевого сигнала.

Пусть определена СММ $Л = (A,B,П)$, тогда вероятность наблюдения последовательности $0 = o_v..o_T$ определяется как

$$P(0\backslash Л) = \quad £ \quad mm\hat{}oM\hat{}M\hat{}-\hat{}г-ЛЛ^{o}\hat{} \qquad (1.5)$$

Вычисление (1.5) путем перебора всех возможных последовательностей неэффективно, поэтому используются специальные методы рекурсивного вычисления, например, алгоритмы прямого и обратного хода [32].

Обучение СММ заключается в нахождении значений параметров СММ, удовлетворяющих некоторому критерию оптимизации, например, максимального правдоподобия (ML, maximum likelihood) [57] или максимальной взаимной информации (MMI, maximum mutual information) [53]. Наиболее часто используется критерий максимального правдоподобия

$$Л^* = \text{argmax}]\text{SogP}(0, | Л) , \qquad (1.6.)$$

где $0_J,...,O_{li}$ - набор обучающих данных для модели $Л$. Аналитического метода решения (1.6), в общем случае нет, поэтому на практике для нахождения параметров $Л$ используется итерационная процедура Баума-Уелча (Baum-Welch) [12], с помощью которой определяется локальный максимум функции правдоподобия (1.6).

1.1.3 Моделирование языковых ограничений

Модель языка служит для описания пространства всех допустимых гипотетических предложений и оценки вероятности *P(S)* каждого

предложения. Двумя наиболее распространенными подходами к моделированию языка в системах распознавания речи являются формальные грамматики [9, 74], использование которых также регламентировано международными соглашениями [34] и статистические п-граммы [36, 52].

Для задач, где все грамматические и синтаксические ограничения или возможные комбинации слов определены, языковая модель обычно моделируется с помощью формальной грамматики. Правила строятся экспертами. Вероятность предложения считается равным 1 если для него существует вывод, в противном случае вероятность полагается равной 0.

Статистическая грамматика основана на оценке вероятности предложения *S*, состоящего из *N* слов *S* -$W_V.W_N$, в соответствии с формулой

$$P(S) = P\{W_V..W_N) = \Pi P(Щ \mid W_V.W_„)$$

Оценка условной вероятности слова от всех предшествующих слов затруднительна. Также очевидно, что вероятность текущего слова в большей степени обусловлена непосредственно предшествующими ему словами. Поэтому обычно используется аппроксимация, что вероятность появления очередного слова в предложении зависит только от предыдущих *n* -1 слов

$$P\{W_t\backslash W_v.W_{i\cdot x}) = P\{W_t\backslash W_{t\_n+v}..W_{t\_x})$$

Такая грамматика называется п-граммой [37]. На практике используются модели со значениями п=1, 2 и 3 (униграммы, биграммы и триграммы). Основным достоинством данных классов моделей оказывается возможность оценки их параметров по реально существующим текстовым корпусам

данных. Условная вероятность $P(JV_i \setminus W_{l\_n+v}..JV^{\wedge}{}_l)$ вычисляется в соответствии с формулой

$$m^{\wedge}\text{-}n+r\text{-}k\text{-}o \qquad ПЩ\text{-}_{11+,}{}^{\wedge}{}_{\_x)} \quad '$$

где F - количество появления данной цепочки слов в обучающем корпусе данных.

1.1.4 Декодирование речевого сигнала

Декодирование речевого сигнала заключается в поиске цепочки слов S^* из множества допустимых цепочек слов языка L, имеющей максимальную апостериорную вероятность $P(S^* \setminus X)$ для заданной последовательности акустических векторов признаков $X = (X_X,\ldots,X_m)$. Согласно формуле (1.4) имеем

$$S^* = \text{argmax}P(S|X) = \text{argmax}P(X|S)P(S) \qquad n_{7)}$$

В силу того, что для каждой цепочки слов S мы можем строить скрытую марковскую модель A_s, формула (1.7) преобразуется в виде

$$S^*=\underset{SeL}{aYgmax}P(X\backslash A_s)P(S)$$

$$\underset{SeL}{\text{argmax}}\left|_{\backslash Q=<1I\text{-}1T}\right| \qquad P(S)$$

Используя алгоритма прямого или обратного хода можно эффективно вычислять $P(X\backslash A_S)$ и, следовательно, вычислять $P(X\backslash A_S)P(S)$ для каждой

гипотетической цепочки слов S.

Проблема при декодировании состоит в том, что множество допустимых цепочек слов языка L огромно и поиск путем полного перебора не представляется возможным. Одним из решений этой проблемы является использование аппроксимации Витерби:

$$S^* = \text{argmax}(\max л: b\{X_x)a \quad b \quad \{X_2)...a \quad (X_T)]P(S)$$

В работе [20] экспериментально показано, что использование аппроксимации Витерби при декодировании не приводит к ухудшению характеристик системы распознавания.

1.2 Методы оценки достоверности результатов распознавания

Как было показано во Введении, существует по меньшей мере две важные проблемы, необходимость практического решения которых приводит к поиску оценок достоверности результатов автоматического распознавания речи. Первая проблема связана с необходимостью идентификации в речевом потоке новых, не входящих в словарь системы (OOV, out of vocabulary) слов или иных акустических событий, а вторая проблема - с возможностью повышения вероятности правильного распознавания за счет использования новой информации, которая может быть получена из каких-то дополнительных источников информации вне системы распознавания речи.

Оценкой достоверности результата распознавания (например, слова) называется числовая характеристика $Cm(JV,X_{iV})$ (обычно между 0 и 1), которая формируется системой распознавания речи для каждого слова W и соответствующей последовательности векторов признаков X_w. При анализе результата распознавания оценка достоверности сравнивается с некоторым порогом T_W . Если её значение больше порога, то слово считается правильно распознанным. В противном случае последовательность векторов признаков X_w считается шумом или незнакомым словом.

Эффективность оценок достоверности результатов распознавания оценивается в терминах ошибок первого и второго вида. Для тестового набора, каждое слово W которого принадлежит либо множеству корректно распознанных *Cor,* либо множеству некорректно распознанных *Inc,* частота

ошибок первого и второго рода, в зависимости от значения порога г, определяются формулами

$$Ещ\{m) \{W \mid W \in Cor\ лCm(Ж,X№) < m\} \mid$$

$$\backslash Cог\backslash \{W \mid W \in Inc\ л$$

$$Cm\{W,Xw) > {}^{m}) \text{ I}$$

$$\text{ii} > r_2(m)$$

Для наглядного отображения эффективности оценки достоверности в зависимости от значения порога используют - операционные характеристики приемника (ROC - receiver operating characteristic curve) [22,26] и характеристики соотношения ошибок обнаружения (DET - detection error trade-off curve) [48,76]. Кривая ROC представляет собой график, где по оси абсцисс откладывается частота правильных принятий *(\-Ггг$_x$(m)),* а по оси ординат откладывается частота пропусков *(Ггг$_2$(m)).* В отличие от кривой ROC, кривая DET описывает соотношение между частотой ложных тревог (*Ггг^(m))* и частотой пропусков *(Ггг$_2$(m)),* взятых в линейном или логарифмическом масштабе.

Кроме графического представления эффективности оценок достоверности, широко используются также скалярные показатели:

- равная частота ошибок первого и второго рода [68,73]

$$EER = Егц\ (г) = Ггг_2\ (г)$$

- минимум суммы частот ошибок первого и второго рода [59,63]

$$TER = \min_{m} (En\backslash\ (г) + En\%\ (г))$$

- общая частота ошибок [46,75]

$$CER = \min_{m} \frac{\backslash Cor\backslash Err_x(z) + \backslash Inc\backslash Err_2(z)}{|\ Cor\ | + 1\ Inc\ \text{I}}$$

Оценки достоверности можно условно разделить на три группы: элементарные оценки, оценки на основе апостериорных вероятностей и оценки на основе отношения правдоподобия.

1.2.1 Элементарные оценки достоверности

К элементарным оценкам достоверности распознаваемого слова относится любая числовая характеристика, получаемая в процессе декодирования. Эти характеристики могут иметь акустическую или грамматическую природу [13, 17, 18, 64, 65]. В качестве признаков для проверки корректности распознанного слова берутся такие характеристики, у которых функция распределения вероятности для правильно распознанных слов существенно отличается от функции распределения вероятности для неправильно распознанных слов. Примеры таких характеристик

- Нормированная акустическая оценка $Cm_{NAS}(W,X_w)$ [16, 38].

Вероятность $P(X_{fV}\backslash A_{jr})$ характеризует степень принадлежности последовательности векторов признаков X_w к множеству акустических событий, которые описываются моделью *Xyy.* При достаточно большом значении $P\{X_W\backslash A_W)$ можно ожидать, что последовательность векторов признаков X_w распознана корректно. Для того чтобы оценка не зависела от длины последовательности X_w, используется её нормированное значение

$$Cm_{NAS}\{W,X_w) = \text{-i-logP(X}_{\text{„}}\text{ I }\backslash_v) \,,$$
$$\text{-V}$$

где W — распознанное слово, X_w — последовательность векторов признаков соответствующей слова W, T_w - длина последовательности X_w, $Я_{,y}$ - акустическая модель слова W. На рис. 3 показана ROC-кривая для нормированной акустической оценки на корпусе данных Ti-digits [55].

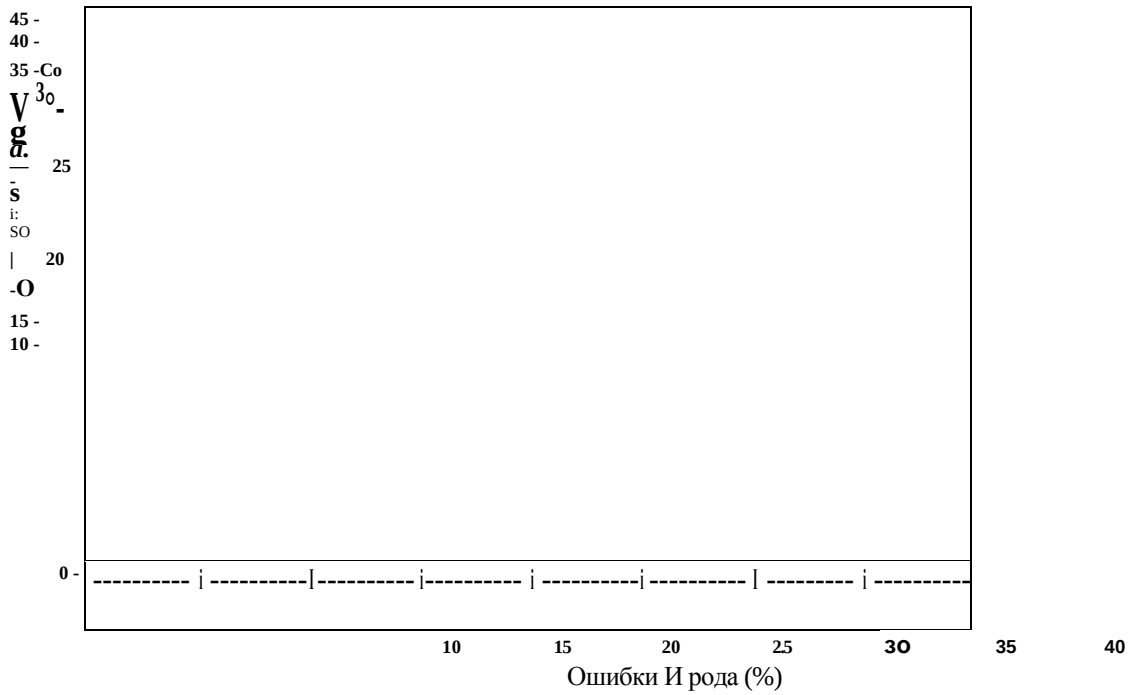


Рисунок 3. ROC-кривая характеристики нормированной акустической оценки

- Плотность гипотез [41]. В процессе декодирования маловероятные гипотезы отсекаются. Поэтому если в некотором сегменте времени вероятность гипотетического слова *W* намного больше чем вероятностей других слов, то большинство из них отбрасывается. С другой стороны, если гипотетические слова имеют близкие и высокие значения вероятности, то отсекание не происходит. Чем больше количество возможных гипотезированных и не отброшенных слов в некотором промежутке времени, тем больше вероятность ошибки в распознавании на этом промежутке. Для каждого слова *W* и момента времени / определяется число гипотез

$$D\{W,t) = \backslash\{W:(W,s,e)eWGAS<t<e\}\backslash ,$$

где *WG* - словный граф (Word Graph), получаемый после процесса декодирования, *s,e* - начало и конец сегмента сигнала для гипотезированного слова *W* соответственно. Тогда для слова *W* плотность гипотез определяется как

$$e\text{-}s + l_{t=s}$$

Численные результаты применения оценки достоверности результатов распознавания в виде плотности гипотез на различных корпусах данных представлены в таблице 2 [78].

Таблица 2

Эффективность оценки достоверности результатов распознавания в виде плотности гипотез.

Корпус данных	Модель языка	Показатель CER(%)
Verbmobil	биграмм	16.6
Verbmobil	триграмм	15.4
Nab	биграмм	14.1
Nab	триграмм	12.5
ARISE	биграмм	10.6
ARISE	триграмм	10.4

Для достижения более хорошего результата применяется комбинация нескольких, взаимно независимых характеристик. Для комбинирования характеристик обычно используются линейный дискриминантами анализ [71, 72], метод опорных векторов [84], нейронные сети [49], дерево принятия решений [23, 51] и т.п. Применяются и более простые методы комбинирования, например использование среднего геометрического взвешенного [55]

$$Cm(W,X_w) = \exp(o,\backslash ogCm,\{W,X_w) + \ldots + a,,\backslash o\hat{}Cm_n\{W,X_w))\ ,$$

где $Cm_x\{W\backslash X_w),...,Cm_n(JV\ ,X_{JV})$ - простые характеристики достоверности слова W, $a_i>0(1<i<n)$ - коэффициенты, удовлетворяющие условию a_x +... + a_n =1.

Методы, основанные на вычислении простых характеристик, просты и не требуют больших вычислительных и временных ресурсов. В то же время, во многих экспериментах было показано, что элементарные характеристики обладают высокой корреляционной зависимостью [39, 41, 65]. Поэтому комбинирование таких оценок часто не приводит к заметному повышению эффективности, по сравнению с использованием характеристик по отдельности. N

1.2.2 Оценки достоверности на основе вычисления апостериорных вероятностей

Вероятностный подход к решению проблемы распознавания речи основывается на теореме Байеса:

$$S_c^* = \underset{Set}{\operatorname{argmax}} \frac{P\{X \mid S)P\{S)}{P\{X)},$$

где $P(S)$ - вероятность модели языки, $P(X \mid S)$ - вероятность акустической модели, $P(X)$ - вероятность наблюдения последовательности векторов признаков X. Если все 3 вероятности известны, то апостериорная вероятность $P(W,s,e\backslash X)$ для конкретного слова W, с началом и концом в моменты времени s u e соответственно, вычисляется по формуле

$$Sel;(tV,s,e)eS \qquad \Gamma\backslash\Pi)$$

Эта апостериорная вероятность могла бы непосредственно использоваться как характеристика для определения корректности распознавания слова W. Теоретически вероятность $P(X)$ имеет вид

$$P(X) = \hat{}_{л} P(X\backslash Л)P(Я) ,$$

где суммирование берется по всем акустическим моделям.

На практике невозможно оценить точно значение вероятности *P(X)* и её рассматривают как величину, которая не зависит от выбора конкретной цепочки слов. Таким образом, решения, принимаемые в процессе декодирования, базируются на ненормированных оценках. Эти оценки пригодны для сравнения конкурирующих цепочек слов, но не для проверки корректности распознавания каждого слова в цепочке. Имеются несколько алгоритмов, которые аппроксимационным образом вычисляют значение *P(X)* с помощью списка N лучших гипотез (N-best List) или словного графа (Word Graph) [60, 69, 78, 79, 81]. Пример при использовании словного графа [80]:

$$P(X) = \sum_{SeWG} P(X \mid S)P(S)$$

и апостериорная вероятность *P(W,s,e\ X)* имеет вид

$$P(W,s,e\backslash X) = \frac{\sum_{SeWG\cdot(W,s,e)\&S} P(X\backslash S)P(S)}{\sum_{SsWG} P(X \mid S)P(S)}$$

где WG - словный граф, получаемый после процесса декодирования. Оценка достоверности для гипотезы *(W,s,e)* вычисляется согласно одной из следующих формул

$$Cm_{N0R}(W,s,e) = P(W,s,e\backslash X)$$

$$Cm_{SEC}(W,s,e) = \sum_{\substack{(W,s\backslash_e y \\ (л,e)п(i',e')\#0}} P<W,s\backslash e'\backslash X)$$

$$Cm_{MED}(W,s,e) = \sum_{s'<(s+e)l2<e'} P(W,sW\backslash X)$$

$$Cm_{MAX}\{W,s,e) = mvL \quad Y \quad P(W,s\backslash e'\backslash X)$$

$$s'<t<e'$$

В таблице 3 приведены значения показателя эффективности CER для оценок достоверности Cm_{NOR}, Cm_{SEC}, Cm_{MED} и $Cm_{ших}$ [80].

Таблица 3

Характеристики оценок на основе апостериорной вероятности

Корпус данных	Показатель CER(%)			
	^{m}NOR	Cm_{SEC}	$^{m}МЕП$	$^{Cm}млх$
NAB20k	10.3	9.2	*9.2*	9.2
NAB64k	8.4	7.2	7.2	7.2
Broadcast News	23.7	20.6	20.4	20.6

Для применения методов, которые используют апостериорные вероятности для вычисления оценки правдоподобия необходимо построение словного графа или списка N лучших гипотез. При большом словаре построение словного графа или списка N лучших гипотез обычно приводит к большому объему вычислений и низкой производительности системы. Кроме того, экспериментально показана зависимость эффективности оценок достоверности от плотности словного графа (отношение количества ребер словного графа к количеству произнесенных слов) и количества гипотез списка N лучших [25, 77].

1.2.3 Оценки достоверности на основе формирования отношения правдоподобия

Данный подход предлагает рассматривать задачу оценки достоверности результата распознавания как проблему проверки гипотез [59, 62, 72]. Пусть имеются распознанное слово W и соответствующая ему последовательность векторов признаков X_w, тогда существуют 2 гипотезы:

H_0 (нулевая гипотеза): последовательность векторов признаков X_w является реализацией слова *W*.

//, (альтернативная гипотеза): последовательность векторов признаков X_w не является реализацией слова *W* и был некорректно распознан как слово *W*.

и отношение правдоподобия:

Если значение $LR\{W,X_W\}$ больше значения порога *m,* то принимается гипотеза H_0, в противном случае принимается гипотеза/^. Таким образом, при известных вероятностях $P\{X_W \backslash H_0\}$ и $P\{X_Ц, |//,)$ мы сможем определить, является ли слово *W* на выходе из распознавателя корректно распознанным.

Чтобы использовать решение на основе (1.8) для каждого слова *W* из словаря системы строятся 2 акустические модели: *Я%* (целевая модель) и *Я^* (альтернативная модель) такие, что $P(X_И,\backslash H_0) = P(X_{ИГ}\backslash X£)$ и $P\{X_W\backslash H^)\text{-}P\{X_{Jу}\backslash Я^)$ для любой последовательности векторов признаков X_w. Если такие модели удалось построить (и оценить их параметры), то в качестве оценки достоверности для слова *W* можно взять функцию

$$Cm(W\,X\,) = —\log^{\wedge A \wedge \wedge '}\,)$$

где X_w - последовательность векторов признаков, распознанная как слово *W* , T_w - длина последовательности векторов признаков X_n,.

В большинство случаев на практике создаются целевые и альтернативные модели не для отдельных слов, а для частей слов (монофонов, Трифонов), из которых составляются все слова словаря системы распознавания [15, 44, 50]. В этом случае, для слова $W\text{-}n_v..u_N$ оценка

правдоподобия вычисляется как функция оценок достоверности составляющих частей, вычисляемых согласно формуле

$$Cm(u,,X_u) = \frac{1}{\log} \qquad \frac{P(X_u \mid Я_н^c)}{} \quad ,\backslash < i < N \ ,$$

где *и,* - часть слова *W, X_u* - соответствующая *щ* последовательность векторов признаков, T_u - длина последовательности векторов признаков X_u, $Я_м^г$ и *Л** - целевая и альтернативная модели для *$u_;$*.

Обучение целевых и альтернативных моделей состоит в нахождении значений параметров моделей так, чтобы они удовлетворяли некоторому критерию. Это может быть, например, критерий максимального правдоподобия (ML, maximum likelihood) [44, 45]

$$(A,f,\wedge) = \text{argmax} \quad (\quad]Г \quad \log P(XJ\wedge) + \quad £ \quad \log P(XJ\wedge) \quad , \qquad \backslash$$

где $X_J,...,X_K$ - набор обучаемых данных, *S{X_k)* - функция, определяемая как

Г1, если $X_к$ корректно распознана [-1,
еслиХд. некорректно распознана

Другим критерием, которым часто пользуется, является критерий минимальной ошибки классификации (MCE, minimum classification error) [59, 67]:

$$\blacksquare \_$$
$$[\wedge 1 + ещ\{-3(X_к)a\{\backslash o\%P\{X_к\backslash ЛЇ)-\backslash_0 £P\{X_к\backslash ЛЇ))- (3$$

$$г$$

$$(4\wedge) = \text{argmin}$$
$$\text{oM})$$

где *а > 0,/3* - числовые параметры.

В работе [45] предлагается построить целевые и альтернативные модели для каждого отдельного трифона *и :* $Я_{г(}^c$ и $Л/^1$, соответственно. Кроме того, имеется одна, т.н. "фоновая", модель $/Г^я$,которая обучается на шумах. Все

модели представляют собой СММ из 3 состояний, эмиссия каждого состояния является смесью из K_c, K_A и $K_в$ нормальных распределений. Величина достоверности трифона *и* для соответствующей ему последовательности векторов признаков X_u определяется как

$$Cm(u,X_u) = \frac{}{T_{Xu}} \log \frac{P\{*u\backslash ti)}{{}^{ь}0.5(P(X_u\backslash Я_u^A) + P(X_u\backslash Л:))\ '}$$

где T_x - длина последовательности векторов признаков X_u. Для слова *W,* состоящего из *N* Трифонов, оценка достоверности вычисляется как

$$Cm\{W,X_w) = \text{^-}f_i \frac{X}{Nttl + exp(-0.5Cm(u_n,X_{Un}))}$$

В таблице 4 приведены результаты экспериментов с разными количествами нормальных распределений при максимального правдоподобия (ML) и минимальной ошибки классификации (MCE) критериях обучения [45].

Таблица 4 Оценки эффективности при минимальной суммы частот ошибок I и II рода (%)

Критерий обучения	K_c = 32	K_c = 32 K_A=Z $K_в$=32	K_c = 32 K_A=4 $K_в$ = 16
ML	28.1	28.3	30.8
MCE *(a = 0.5,j3 = 0)*	26.8	27.2	27.1

Главная проблема методов, основанных на отношении правдоподобия, заключается в удачном выборе и моделировании альтернативных моделей. Это объясняется тем, что пространство акустических событий, которые

должны моделироваться альтернативными моделями очень большое и сложное. Кроме того, методы обучения целевых и альтернативных моделей так же играют важную роль в эффективности метода.

1.3 Выводы

Приведен обзор вероятностного подхода к моделированию и распознаванию речи. Показано, что применение вероятностного подхода требует построить четыре составляющие модули: извлечения векторов признаков сигнала, акустико-фонетического моделирования, моделирования языка и декодирования. Для каждого составляющего модуля системы распознавания речи подробно описаны их функции и способы их построения. Дано описание скрытой Марковской модели, которая широко используется для построения акустических моделей. Приведены критерии обучения СММ. Приведено описание метода декодирования, основанного на использовании алгоритма Витерби.

Рассмотрен графический способ предсавления эффективности оценок достоверности, который состоит в построении характеристик ROC или DET. Также описаны скалярные показатели эффективности оценок достоверности, такие как расвная частота ошибок первого и второго рода, минимум суммы частот ошибок первого и второго рода, общая частота ошибок.

Проведен анализ основных подходов к формированию оценки достоверности для систем распознавания речи. Оценки достоверности предложено условно разделить на три группы: элементарные оценки, оценки на основе апостериорных вероятностей и оценки на основе отношения правдоподобия. Указаны недостатки каждого подхода.

Показано, что подход, основанный на построении отношения правдоподобия представляется наиболее перспективным для решения проблемы синтеза оценок достоверности. В отличие от других подходов, данный подход менее зависим от параметров системы распознавания и не

требует полного отслеживания процесса декодирования. Использование этого подхода, в свою очередь, предполагает разработку метода построения целевых и альтернативных моделей и алгоритма обучение этих моделей.

Глава 2. Оценки достоверности на основе отношения правдоподобия

Глава содержит описание новых методов формирования оценок достоверности, которые относятся к методам на основе отношения правдоподобия. В отличие от существующих методов данного класса, где отношения правдоподобия вычисляются для каждого слова или части слова, в данной работе предлагается использовать значения отношения правдоподобия на уровне отдельных векторов признаков.

2.1 Выбор моделей для построения отношения правдоподобия

Пусть дана система распознавания речи, основанная на вероятностном подходе с использованием СММ. Тогда для последовательности векторов признаков $X = (x_x,...,x_m)$, распознанной как слово W, *мы* можем однозначно найти оптимальную последовательность состояний СММ $Q = (q_x,..;q_T)$ в соответствии с соотношением

$$Q = \underset{(\text{ïï.-..,?/-})}{zrgmaxP(X\backslash(q_b...,q_T),A)}$$

$$= \underset{(</i>\ .</i/)}{\text{argmax}}(\wedge_i\wedge_|(x_1)\wedge_{i(/2}\wedge_2(x_2)...{}_{Vi?7}.\wedge_7.(x_r))'$$

где $Л$ - параметры СММ для слова W, *ж* - вероятность начального состояния q_x, $b_q\{x)$ вероятность появления вектора признаков x в состоянии $q.$, a_q - вероятность перехода от состояния q_{t_x} к состоянию $q.$. Таким образом, каждый вектор признаков x_t ассоциируется с некоторым состоянием q_t.

Для каждой пары (x_nq_t) построим 2 гипотезы:

H_0 (нулевая гипотеза): последовательность векторов признаков X корректно распознана как слово W при данной паре (x_nq_t).

H_x (альтернативная гипотеза): последовательность векторов признаков X некорректно распознана как слов W при данной паре (x_t,q_t).

Согласно правилу принятия решения по максимуму апостериорной вероятности, принимается нулевая гипотеза H_0, если $P(H_0 \backslash x_t, q_t) > P\{H_X \mid x_t,q_t)$. В противном случае, принимается альтернативная гипотеза Я,. Это правило можно записать в следующем виде

$$\text{принимается гипотеза } H_0, \text{ если } \frac{P(H_0\backslash x_t,q,)}{P(H_0\backslash x_{lt}q_t)} > 1$$

$$\text{принимается гипотеза } H,, \text{ если } \frac{P(H_0\backslash x_t,q_t)}{} < 1 \qquad (2.1)$$

Применим формулу Байеса к вероятностям $P(H_0 \backslash x_t, q_t)$ и $P\{H_X \backslash x_n q_t)$ в (2.1)

$$P(x_t,q_t)$$

$$P(x_t,q_t)$$

В результате получим

$$\text{принимается гипотеза} H_0, \text{ если } LR(x_t,q_t) > m^y$$

$$\text{принимается гипотеза} H_x, \text{ если } LR(x_t, q_t) > z^J$$

где

$$LR(x_n q,) = \frac{P(X \backslash H_0 \backslash t,4,J)}{P(x_t,q_t\backslash H_t)}$$

$$P\{H_0)$$

Так как значения вероятностей $P(H_0)$ и $P(H_X)$ неизвестны, то значения порога m' выбирается эмпирическим образом в зависимости от состояния q_t, т.е. г' = г' .

Определим элементарную функцию достоверности на уровне вектора признаков

$$C(x_b q,): \quad \frac{1}{l + \quad ^{ч-1}(LR(x_b q,))-}$$

Элементарная функция достоверности обладает следующими свойствами:

1. Если $C\{x_t, q_t) > r_{qi} = 1 / (1 + (г')^{-1})$, то принимается гипотеза $Я_0$. В противном случае, принимается гипотеза H_x.
2. Для любой пары (x_t, q_t) выполняется условие $0 < C\{x_t, q_t) < 1$.

Вычисление значения элементарной функции достоверности $C(x,q)$ требует определения распределений $P(x,q\backslash H_Q)$ и $P(x,q\backslash H_l)$ для каждого состояния q. Определим модели $Ф_д$ и ¥ таким образом

$$P(x,_д\backslash H_0) = P(x\backslash Ф_д)$$
$$P(x,q\backslash H_x) = P\{x\backslash 4>_q)$$

Модели $Ф_?$ и $Ш_q$ будем называть целевой и альтернативной моделями для состояния q соответственно. Предполагаем, что распределения $P(x\backslash Ф)$ и $P(x \mid *F)$ являются смесями нормальных распределений

$$_{м_0} P(x \mid Ф,) = Y, c_{Q>k} N\{x, m_{QJo} v_{Qk})$$
$$к=\backslash$$

$$к=\backslash$$

где $c_{0к}>0$ и $c_{1к}>0$ - веса нормальных распределений, удовлетворяющих условиям

$$_{A=1}^{м_n} 2\text{-}i^{C}0,k \sim _{A=1}^{м,} 2\text{-}1^{C}\backslash I \sim *\blacksquare \quad '$$

N(x,m,v) - нормальное распределение со средним $m = (m_{(1)},...,w_{(D)})$ и дисперсией $v = (v_{(1)},...,v_{(D)})$

$$N(x,m,v) = \frac{1}{\prod_{y=1}^{*} V'' * 2\text{-}\Gamma \; v_w} \text{ з-exp } -X,$$

D - размерность векторов признаков.

Таким образом, для каждого состояния скрытой Марковской модели звука мы определили целевую и альтернативную модели, а также построили элементарную функцию достоверности для вектора признаков.

2.2 Методы формирования оценок достоверности

На основе значений элементарной функции достоверности $C\{x_t,q_t)$ для каждого вектора признаков x_t последовательности $X = (x_l,...,x_T)$, распознанной как слово *W,* строим оценку достоверности. Исходя из представленного способа построения и свойств функции $C(x_t,q_t)$, можно предположить, что последовательность векторов признаков *X* корректно распознана как слово *W,* если

$$C(x_t,q_t) > T_{qi}, \backslash < t < T \qquad (2.2)$$

Тогда в качестве оценки достоверности распознанного слова *W* можно взять функцию

$$Cm(W,X) = \backslash \begin{cases} 1, \text{ если VI} < t < T: C(x.,q.) > r_{a_q} \\ \text{[O, если 31} < t < T: C(x_t,q_t) < r_{q\%} \end{cases}$$

Следовательно, при анализе результатов распознавания, последовательность векторов признаков *X* считается корректно распознанной как слово *W* если $Cm(JV,X) = \backslash$. В противном случае, $Cm(W,X) = Q$, последовательность векторов признаков *X* считается некорректно распознанной.

Условие (2.2), при выполнении которых последовательность векторов признаков *X* считается корректно распознанной, являются очень сильными ограничениями. На практике неизбежны ошибки, когда последовательность векторов признаков *X* корректно распознана как слово *W*, но имеются такие пары (x_t,q_t), что

$$C(x_t,q_t)<T_{qt}$$

Прріпимая это во внимание, предлагается формировать оценку достоверности *Cm(W,X)* на основе средних значений, одним из следующих способов:

$$Cm\{W_9X)\hat{}a_tC\{x_{ti}q_t) \qquad (2.3)$$

$$Cm(W,X) = exv\ £a,\ \mathrm{In}\ C(*„?,) \quad , \qquad (2.4)$$

где $a_t > 0(1 < t < T)$ - весовые коэффициенты. Чтобы оценки достоверности (2.3) и (2.4) не зависели от длительности *T,* наложим ограничения на значения весовых коэффициентов

Предложенные таким образом методы формирования оценок достоверности для слова будем называть одноуровневыми методами.

Последовательность векторов признаков *X* будем считать корректно распознанной как слово *W,* если значение *Cm(W,X)* больше чем значение порога T_W. В противном случае, $Cm(W,X)<T_w$, последовательность векторов признаков *X* считается некорректно распознанной как слов *W*. Значения порога $m_ц$, является фиксированным и заранее выбирается, например, эмпирическом образом для каждого слова *W* словаря системы

распознавания речи. Кроме того, из войства элементарной функции достоверности следует, что

$$0<Cm(W,X)<l$$

2.2.1 Двухуровневый метод формирования оценок достоверности.

В разделе 1.1.2 показано, что для системы распознавания речи с большим объемом словаря акустические модели, как правило, строятся для контексто-зависимых реализаций фонем или фонов. Поэтому кроме распознанного слова *W,* на выходе из декодера также определена соответствующая ему последовательность образующих фонов.

Пусть распознанное слово *W* состоит из *N* фонов, т.е. W -$u_x u_c ... u_n$. Для фона $u_n(\backslash<n<N)$ обозначим соответствующие ему последовательности векторов признаков и оптимальную последовательность состояний через $^{X}u_n=\{x_{Sn},x_{Sn+}x>->^{x}e)$ и $Qu_n=\wedge_{Sn},q_{Sn+x},-,q_c)$, соответственно. Последовательности X_u и $Q_{„}$ являются подпоследовательностями $^{"и}$ $^{"и}$ последовательностей X и Q и

$$s_x=\backslash, \quad e_N=T\ s_n = e_{n\_x}+\backslash,2\wedge n<N$$

Аналогично (2.3) и (2.4) определим оценки достоверности $Cm(u_n,X_u)$ для фона $u_{,,}$ и последовательности векторов признаков $X_{,,} = (x\wedge,x, ,,,...,д;)$: $^{°n}$ ^{J_n}

$$Cm\{u_n,X_u)\wedge Y,a_tC\{x_nq\wedge \quad (2.5)$$

$$Cm(u_n,X_u) = \exp \left| \frac{£\wedge,\ln C(x_{„},?_r)}{4'=*.} \right| \quad (2.6)$$

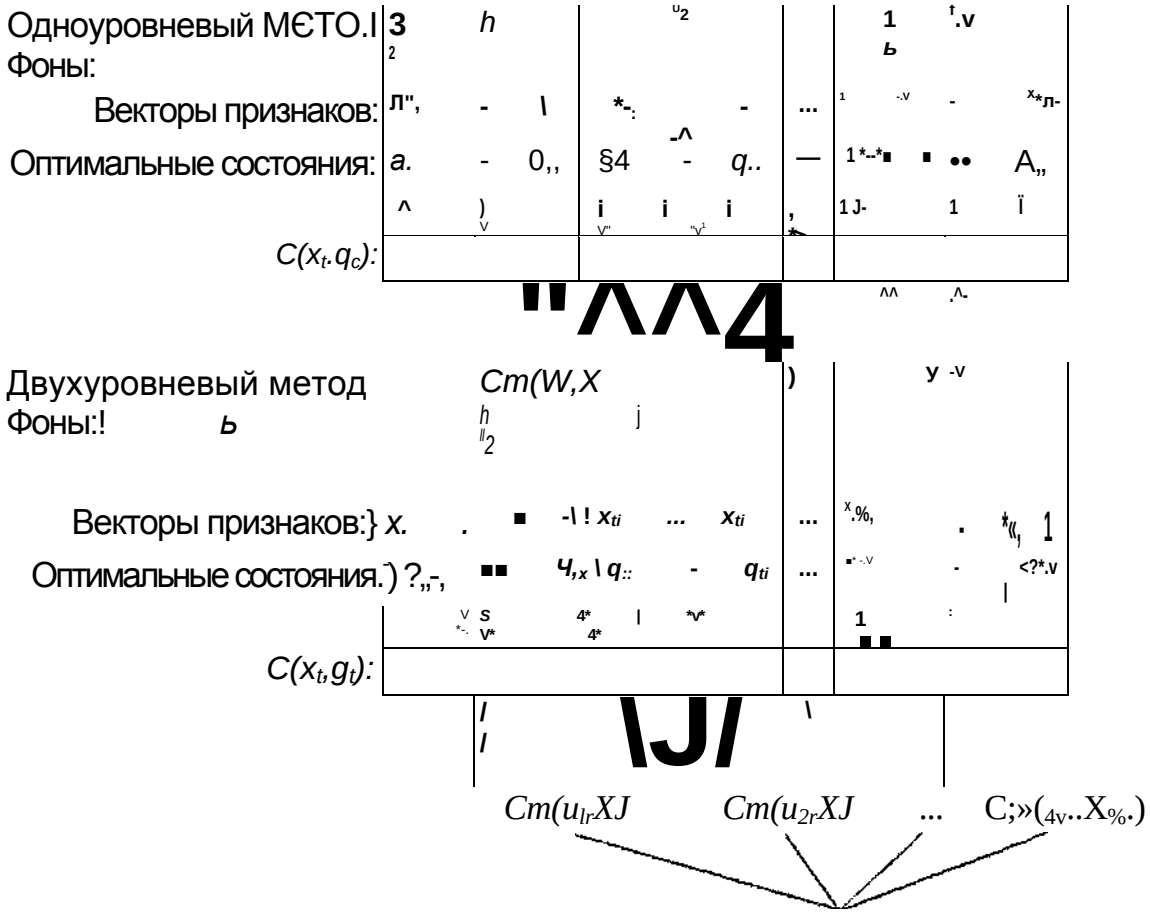
где $a_t > 0(\wedge,, < t < e_n)$ - весовые коэффициенты $u\ a_{s\ "n\ "д'} + a_{s+1}+... + a_e = 1$

Для слова *W* формируем оценку достоверности как арифметическое среднее или геометрическое среднее оценок достоверности для составляющих его фонов:

$$Cm(1¥,X) = \underset{M}{\wedge} Cm(u_n,X_{un}) \qquad (2.7)$$

$$Cm(W, X) = \exp \frac{(1}{V\wedge} \pounds \underset{л=1}{\overset{'V}{\wedge}} \ln Cm(u_н, X_{Vf}) \qquad (2.8)$$

Методы формирования оценок достоверности согласно формулам (2.7) и (2.8) будем назвать двухуровневыми методами. Различие между одноуровневыми и двухуровневыми методами формирования оценок достоверности для слов наглядно представлено на следующей рисунке.



Єж(Г,Х) Рисунок 4.

Схемы формирования оценок достоверности

Отметим, что использование комбинация формул (2.5), (2.7) или (2.6), (2.8) для формирования оценок достоверности слов, по сути, является частным случаем (2.3) или (2.4), соответственно. Действительно:

$$Cm(W,X) = \pm\text{-}JTCm(u_n,XJ$$

$$iVf, = £дс(*„<7,)$$

$$Cm\{W,X) = exp\backslash\pm f\backslash nCm(u_n,X_{Un})$$

$$= \exp \quad \exp \quad \exp \quad \text{ХД1пС(х,,^)}_J$$

где

$$A = \begin{cases} \frac{1,/}{N} \text{ or } , s_x < t < e_x \\ \frac{Nj}{a_N} , s_N < t < e_N \end{cases}$$

$j3_(> 0(1 < t < T)$ и удовлетворяют условию

2.2.2 Задание весовых коэффициентов

Пусть имеется выборка из K последовательностей векторов признаков $\{X^{>)}\}(1<1<K)$, где $X^{(l)} = \{x\backslash'^],...,x^{\wedge})$, и каждая из последовательностей векторов признаков $X^{0)}$ распознана декодером как слово $W^{(,)}$ с

соответствующей оптимальной последовательностью состояний $Q^{(l)}$ -(Я\\—,4^)- Выборка $\{X^{i0}\}$ считается достаточно большой и содержит большое количество как корректно распознанных, так и некорректно распознанных последовательностей векторов признаков для каждого слова словаря системы распознавания речи.

Для каждой пары (.^,^)(1 < / < K,\ <m<$T_{(}$) определим функцию

*ttM}H 1, если^Γ(,) корректно распознана как слово $W^{(,)}$ -1, еслиX(,) некорректно распознана как слово $W^{(l)}$

Через C_q и I_q обозначим множества векторов признаков для любого состояния q такие, что

Пусть для состояния q уже построены целевая и альтернативная модели (Ф и W соответственно), введем дискриминационную величину d_q

$$\backslash Mc \sim Mi \mathrm{I}$$

$$<j_r + <7,$$

где

$$u_z = \text{------} \vee C(x, a)$$

$$°c_ч = T77\text{-}\wedge\{^C\wedge^x\wedge\text{-}\wedge c_ч)^2$$

$$Cq \mathrm{I} *€C,$$

$$** = U1^{C\{X4)}$$

$$q\ ix\epsilon/q$$

$$/_{?}=\text{гп}E(^{c}(\wedge)\text{-}\wedge)^{2}$$

Согласно формулам (2.3)-(2.6), для последовательности векторов признаков $X_z = (x_{z<l},\ldots,x_{ZtT})$, которая распознана декодером как единица речи

z (слово или фон) с оптимальной последовательностью состояний $Q\text{-- }=(Qz,\backslash>\blacksquare\blacksquare>{}^{C}Iz,T.)$-> оценка достоверности $Cm(z,X_z)$ определяется как

$$Cm(z,X_z) = \wedge a_{21}C\{x_{24},\partial_{2(}) \quad (\text{г.})$$

$$Cm(z,X_z) = exp\backslash\ Ys^{a}z;^{lnC}(^{x}z,t>q;,<)$$

Предлагается следующий способ задания значений весовых коэффициентов

$$",,= \frac{''^{+}}{2X7},\ 'a\,г<\Gamma_{г},$$

где $p > O$ - числовой параметр.

Очевидно что, предложенные таким образом весовые коэффициенты $a_{zl} > 0(1 < / < T.)$ и удовлетворяют требуемому условию

$$^{a}z,X + \wedge,2+\text{-} + \wedge=1$$

При значении $p = 0,$ оценка достоверности $Cm(z,X,)$ представляет собой арифметическое или геометрическое среднее значений $C(x_{zl},q_{:t})(l<t<T_z)$:

$$Cm\{z,X_z) = \sim f_jC\{x\wedge qJ$$

$$Cm(z,X_z) = exp\backslash\sim j]\backslash nC(x_{zJ},{}_{qzJ})$$

При значении *p->oo,* оценка достоверности *Cm{z,X_z)* имеет один из следующих видов

$$Cm(z,X_z) = \frac{}{A}$$

$$Cw(z,X_z) = \exp\left(\frac{1}{} \sum \ln C \right)$$

где fiU = $\max_{\backslash<t<i.}$^, *A* =| fo, | rf = <шч л 1 < *t* < *T_2)*

2.3 Обучение целевых и альтернативных моделей

Обучение целевых и альтернативных моделей состоит в нахождение значений их параметров согласно некоторому критерию обучению. К числу параметров каждой целевой или альтернативной модели, подлежащих нахождению, относятся размер модели (т.е. количество смесей нормальных распределений), веса смесей, средние и дисперсии смесей. Обучение будем проводить на множествах векторов признаков С и / , которые описаны в разделе 2.2.2. В следующих разделах содержат описания критерия обучения и предлагаемого алгоритма обучения.

2.3.1 Критерий обучения моделей

Пусть для состояния *q* уже построены целевая и альтернативная модели (Ф? и Ч/ $_q$ соответственно) и найдено значение порога *m* . Тогда для данного вектора признаков *x* имеет место ошибка, если

$$\begin{array}{l} C(x,q) < z_g, xeC_g \\ C(x,q) > T_q, xtI_{(/} \end{array} \qquad (2.9)$$

Условие (2.9) эквивалентно условию

$$S(x,q)(r_q - C(x,qj) > 0, xe\{C_q \wedge I_q\}$$

Определим функцию ошибки как сумму частот ошибок при заданных моделях $\Phi_д$, $^x¥_q$ и значении порога m

$$\Pi\Phi_д, У_ч, m_я) = -\wedge\text{-}T8I_e n\{3\{x,я)\{m_ч\text{-}C\{x,я))) \\ — Ysign(S\{x,q)(r_q \text{-C}(*,g))) \tag{2.10}$$

где $sign(z) = 1$ при $z > 0$ и $sign(z) = 0$ при $z < 0$.

Функцию $sign(z)$ можно аппроксимировать с помощью сигмоидальной функции [33]

$$sign(z)*\text{- ------ —}, \quad 1 + e \tag{2.11}$$

где $a > 0$ - числовой параметр.

Применяя (2.11) к (2.10) получим

где

$$R(x,O_q,4>_q,T_q) = \text{----------}_г \frac{1}{1 + exp\backslash a_q S(x,q)\backslash]nLR(x,q) - b_q jj}\text{----------}_п,$$

$a > 0$, Ъ - числовые параметры, выбираемые в зависимости от значения г .

Тогда обучение целевых и альтернативных моделей предлагается проводить отдельно для каждой пары (Ф^,^) таким образом, чтобы сумма

частот ошибок $Г(Ф\ ,4^*\ ,m\ )$ была минимальной, т.е.

$$(O_q,4>_q) = \underset{(\%Y_q)}{argmm} F(O_q,4!_q,T_q)$$

$$\text{argmin}\left|^{I}\right|$$

2.3.2 Обучение моделей методом градиентного спуска

Пусть для моделей Φ_∂ и $^{x}¥_q$ распределения Р(л*|Ф^) и $P(x\backslash 4^{J}_q)$

являются смесями нормальных распределений из M_0 и M, компонентов, где значения M_0 и М, заранее выбраны.

$$\text{^|ф?)=Ec}_0\text{/(x,\%^}_{0i}\text{,)}$$

$$\text{A=1 Л/,}$$

$$\mathbf{P(*|^) = 2>_u №,m_u,v_u)}$$

С помощью метода градиентного спуска можно находить значения параметров моделей Φ_∂ и $Ч^*$, при которых значение функционала ошибки

$F^\wedge_q,^{x}¥ \quad z_q)$ является локальным минимумом. Итеративная формула метода градиентного спуска имеет вид

(0^^^) = (^^^)-^^^^^^,)^ = 0^-- $C^2\text{-}^{12}$) где $\epsilon>0$ - числовой коэффициент, $Ф^{н)}$ и $4^{'\circ}$ - значения параметров

целевой и альтернативной моделей на n-ой итерации, $Ф^{0)}$ и $Ч^{0)}$ - начальные приближения.

Итеративные формулы (2.12) сходятся к локальному минимуму, т.к. функция ошибки $F(<\&_q,^{x}¥_q,T_q)$ является непрерывной и ограниченной снизу $(F(0_q,V_q,T_q)>O)$.

При применении итеративной формулы (2.12) необходимо принять во внимание следующие ограничения на значения параметров (весов смесей и дисперсии) целевой и альтернативной моделей

$$c^{>0,} \overset{M_{„}}{\underset{Ar=|}{IX,}}{}^{*-1}$$

$$v_{ffJcJ} > 0,(cm = 0,l;k = l,2,...,M_a;i = l,2,..,D)$$

Выразим параметры c_{ak},v_{akj} через дополнительные параметры c_a^,^$_a$^. следующим образом

$$\overset{c<m,к}{M,} \quad \mathbf{1}$$

$$\mathbf{Z^{ex}p(^)}$$

$$v^{\wedge} = езф(y_{aii1>1}),((7 = 0,1;* = 1,2,..., M_{ff};i = l,2,..,D)$$

Обозначим через 7^((7 = 0,1) некоторый параметр из {сГ.,$_k$,m_{akl},v_{aJcj}}. Частную производную $dF(<\&,4^*,r_q)\ I\ dz_a$ можно представить в виде

$$\mathbf{dz_{„}} \quad \underset{I\ Q\ I\ *\epsilon C_{„}}{\overset{'q\ i\ \blacksquare"=<-,?}{}} \quad \overset{\&_{„}}{} \quad \underset{xe\{C_{ч}ul_{я},}{} \quad \overset{9\quad I\ A^{\wedge}'q}{} \quad \boldsymbol{dz_{„}}$$

$$= \quad X \quad {}_r(^c7)(l\text{-}i?(x,O_g,^,r_g))^{\wedge}(x,O_g,^,r_g)^{\overset{ai0g}{}\overset{(x|e)}{}}$$

где 0 = Ф$_?$ если (7 = 0 или 0 = 4^ если ¢7 = 1,

$$y(x,cг) = < \begin{cases} a_{iи}/\backslash C_{\partial}\backslash,xeC_{\partial}ла = 0 \\ -a_q/\backslash I_q\backslash,xeI_qA<T = 0 \\ -a_q/\backslash C_q\backslash,xeC_qAa = \backslash \\ a_{\partial}/\backslash I_{ч}\backslash,x\epsilon I_{ч}лcm = 1 \end{cases}$$

Для параметров $c_{ак}$, m_{aki} и v_{aki} частные производные $d\backslash ogP(x\backslash\&)/\partial c_{ак}$. $d\backslash ogP(x$ 10)/ $\partial m_{акi}$ и $d\backslash ogP\{x \mid 0) / dv_{акi}$ вычисляются по формулам

$$\frac{\partial \log P(x\ 10)}{\overset{a,к}{\partial e.}} \quad \frac{c_{aЛ}Щx,\ m_{aЛ},\ v_{aM})}{P(x\backslash@)} \quad {}^{'aji}$$

$$31\text{ogP}(x|0) \_ c_{atk}N(x,m^{\wedge}_{k},v_{aA})\ (Xi\text{-}m^{\wedge}j)$$

$$dm_{a,k,i} \qquad P(x\backslash@) \qquad v._{a,k,i}$$

$$8\log P(x\ 10) \quad c_{aJt}N(x,\ m_{aJt}\ ,Y_{aJ}) \quad (.Xi\text{-}m^{*}jtjY$$

$$2P(x\backslash\epsilon) \qquad {}^{Y}a,k,i$$

$$dv.$$

Тогда для параметров целевой и альтернативной моделей итеративные формулы представляются в виде

$$(\text{я}) \quad \text{ш}(")$$

$$<\#\ \exp \qquad \partial z(\Phi^{\wedge},^{\wedge};\backslash m_{\partial})$$

$$(11+1)\ \_\ \_ \qquad V \qquad \text{-e} \qquad \partial c^{(n)}$$

$$\text{аПФ}^{\wedge}\text{ГЧ})$$

$$A=1 \qquad \text{Эс}^{(\text{п})}$$

$$(") \quad \text{и/(и)}$$

$$(\text{л+1}) \_ \qquad (\text{и})$$

$$'\text{cr.U} \quad \sim^{m}a,k,i$$

$$\circ \qquad \text{с} \quad (\text{я})$$

$$v(\text{«+i}) \_ v(\text{«}) \ \text{exn} \qquad \text{-£•} \qquad \frac{\text{ai?(ф(П)}:\text{УуОО}_{?\text{г}}{}^{\wedge}}{\text{CT,ft,(}} \qquad ,(0\text{-} = 0,1^{\wedge} = 1,2,..,^{\wedge}/ = 1,2,..,1))$$

Описанный метод оптимизации позволяет просто и эффективно находить значения параметров моделей Ф и W с локальным минимумом

^(Ф ,4^{y} ,r). Однако данный метод обладает некоторыми недостатками.

Первым недостатком, свойственным алгоритму градиентного спуска,

является зависимость от начального приближения Ф^ и $*¥_{q}$. Второй

недостаток заключается в том, что необходимо заранее выбрать числа компонентов смесей нормальных распределений для описания распределений *P(x\Ф)* и *P(x\4?$_{q}$).*

2.3.3 Улучшенный алгоритм обучения моделей

В предыдущем разделе приведен алгоритм оптимизации моделей, основанный на методе градиентного спуска, и указаны недостатки

этого

алгоритма. В этом разделе предлагается модифицированный алгоритм, позволяющий решить проблемы выбора размеров моделей и начальных приближений.

Идея алгоритма заключается в следующем. Имея целевую и альтернативную модели, распределения $P(x\backslash\Phi_{\partial})$ и $P(x\backslash^{x}¥\ )$ которых

являются смесями нормальных распределений из M_{ϕ} и M_{Y} компонентов, попытаться увеличить M_{ϕ} или $M_{ч}$, на единицу с целью уменьшения ошибки F(<I> ,^,^). На начальном шаге алгоритма распределения P(x|Ф?) и $P(x \mid Ч')$ описываются однокомпонентными смесями, т.е. $M_{\phi} = My$ -1.

Для реализации алгоритма необходимо определить способ генерирования новой модели 0' из имеющейся модели 0, где распределения $P(x10)$ и $P(x10')$ являются смесями из M и M +\ компонентов, соответственно:

$$\begin{array}{c} {}_{A/}P(x|0) = \\ X<yV(.Y,W,,V,) \\ \kappa\backslash \end{array}$$

$$P(x\backslash e') = \overset{M+\backslash}{Y,c\backslash N(x,m\backslash y_k)}$$

Генератор новой модели зависит от выборки векторов признаков Z и параметра г($1 < г < M$), т.е.

$$0' = Генерирование^, Z, г)$$

Обозначим через $Z_{(г)}$ множество векторов признаков, где

$$Z_{(г)} = \underset{\{}{ïx}\ I\ x \in Z\ л\ \underset{\backslash<\kappa<M}{argmax}[c_{it}A''(x,m_{;t},v_{Ar})] = г\ \backslash \quad)$$

С помощью метода k-средних находим значения векторов $/u_0$ и //,, минимизирующие сумму

$$J(Z_{lr},H>,Mi)= \sum_{\substack{xe \\ Z,}} \mathbf{I}^{x}\text{-}\wedge\text{-} \quad (*)$$

где /ir(jc) = $\underset{1=0,1}{\text{argmin}}$(||x-//,1)

Метод к-средних состоит в обновлении значений $//_0$ и //j согласно формулам

$$//_0 = £\{л; \mid x \in Z_{(r)} \text{ v } лг(л:) = 0\}$$

$$Д = £"\{д: \mid л: \in Z_{(r)} \text{ v } /f\ (x) = 1\}$$

В качестве начальных значений установим

$$//_0 = £\{л: \mid л: \in Z_{(r)}\} - \epsilon$$
$$//, = £\{x|xeZ_{(r)}\} + 6\text{-}$$

где £• - вектор, элементы которого являются достаточно малыми числами.

Зададим начальные значения параметров модели 0' следующим образом

$$(^{c}\backslash>m\backslash У_{к}) = \left| \begin{array}{l} (c_r/2,ju_Q,v_r),k = r \qquad ,1<к<M + \backslash \\ (c,./2,//,,уД£ = M + 1 \end{array} \right.$$

Окончательные значения параметров модели получаются в результате применения алгоритма EM (expectation maximization) с тем, чтобы максимизировать функцию правдоподобия на выборке Z

$$L(Z\backslash@') = Yl^{P}(\backslash®')$$

Итеративные формулы EM алгоритма имеют вид [14]

$$\sum_{\substack{xeZ}} T,PnЛ*)$$

$$\sum_{x \in X} \left(\hat{}(*x*,->«v,:) \right)^2$$

$$\pounds \#..*(*)$$
$$xsZ$$

где

$$AW = P(x\backslash в<)$$

Ниже приводится пошаговое описание предлагаемого алгоритма обучения целевой и альтернативной моделей для состояния q.

Алгоритм оценки параметров моделей.

1. Инициализировать модели $Ф_{Сигг}$ и $^{X}Р_{Сигг}$. Установить

$$М_ф = 1,-М\hat{} = 1\ (Фьс'Ч'ьпг) =$$

$$ГрадиентныйСпуск(Ф_{Сигг}, Ч'_{Сигг})$$

2. Для всех *QJ)* є {(1,0),...,($М_ф$,0),(0,1),...(0,^)} выполнять

$$Ф,\hat{} \quad <*W = 0$$
$$Генерироваше(Ф_{Curr}, C_q, i), i\ Ф\ 0$$

$$П/= \left| \begin{array}{l} Фс\hat{}7 = 0 \\ Генерирование\hat{}_{Curr}\ J_q, j\backslash j\ Ф\ 0 \end{array} \right.$$

$$(Ф?\ .,4\hat{}\ .) = ГрадиентныйСпуск\{Ф_i\ j,\hat{}_{\{}j)$$

3. Установить

$$(Л/) = аг_{ё}тт(\hat{}(Ф\hat{}\hat{},г))$$
$$(\cdot.У)$$

$$ф \quad = ф \quad w \quad = w$$

$$Л\hat{}Ф = M_ф + 1, \text{если } i* Ф\ 0$$

$$М_v = Му + 1, \text{если}./'* \hat{}\ О$$

4. Если Д = ^(Ф$_{ів}$Р¥$_{йв/}$,г)-.Р(Ф^,,¥*.$_{(у,,г)}$ > 0, то установить

$$\text{ф}_{\text{^best}} = \text{ф}^*_{\text{^/*,y*'}^T} \quad w_{best} = w^*_{T\, i^*,j^*}$$

5. Если выполняется критерий останова, то закончить алгоритм с моделями ®*best* и ^*best* Д$^{\text{ля}}$ Целевой и альтернативной модели соответственно. Иначе перейти к шагу 2.

На шаге 1 строятся текущие модели $\Phi_{Сигг}$ и $Ч_{Сигг}$ для целевой и альтернативной моделей соответственно. Модели $\Phi_{Сигг}$ и ^$V_{Сигг}$ состоят из параметров {$c_{ол}$,/%,v_{0J}} и {$c_{1Л}$,/$н_{1Л}$,$т_ц$}, где

$$m_{0J} = \text{------}\,\text{^}\Gamma x, \; m_{1Л} = \text{------}\,\text{^}\Gamma\,\text{x}$$
$$\text{IЦI}\,{}^*cc_ч \qquad\qquad \text{I}U_q\backslash xel_ч$$

$$\frac{1}{\text{y}}\;\text{O, -}^m oxf{>}^v w = \Gamma\Gamma 7\text{Z}(^x i \sim^{\text{T}}\text{i,и})^{2\text{ГГ}'} =$$
$$\text{I Š I}\,\text{^}\epsilon\text{C}, \qquad\qquad \text{I}\,\text{^}$$

Текущие количества компонентов $M_ф$ и М^ смесей распределений $P(x \mid \Phi)$ и Р(л' | Ч') устанавливаются равными единице. Производится оптимизация моделей методом градиентного спуска с текущими моделями $\Phi_{Сигг}$ и ¥$_{Сшт}$ в качестве начальных приближений. Полученные модели сохраняются как лучшие из найденных $\Phi_ш$ и $Ч_{best}$. На шаге 2 строятся $M_ф$ + M^ пар новых моделей (Ф. .,^..), где одна из моделей Ф$_{/у}$ или $Ч^*_и$ получается в результате генерирования новой модели из $\Phi_{Сигг}$ или ^¥$_{Сигг}$ соответственно, а другая не изменяется. Для каждой новой пары моделей (Ф. ,4^. .) вычисляется (Ф*y_5^*,у) ^{c} помощью градиентного спуска. На шаге 3 текущие модели $\Phi_{Сигг}$ и $^V F_{Curr}$ устанавливаются равными Ф.» .* и ^^*j**,* значение ^(Ф,.*^,,^*, ,,т) которых является наименьшим среди новых пар моделей.

Значения $M_ф$ и $M_ч$, устанавливаются равными числу компонент смесей распределений $P(x\backslash Ф_r\wedge)$ и P^l^,* .»). На шаге 4 выполняется проверка: если значение ^(Ф,*j*,^*» *,г) меньше, чем значение ^($Ф_{йет}$,,T\,Л,г), то модели Ф,*$_{>7}$* и $4^{у}$.*$_{у}$* сохраняются как оптимальные. На шаге 5, если выполняется критерий остановки, то алгоритм завершается с <£>$_{best}$ и *¥$_{best}$ как целевой и альтернативной моделями. В противном случае перейти к шагу 2. В качестве критерия останова предлагается использовать следующее условие: A < ϵ и сумма $M_ф$ + M^ больше чем некоторое число M_m-$_n$.

Предложенный улучшенный алгоритм сходится, так как функция ошибки $F(\wedge_q,^xV_q,T_q)$ ограниченна снизу $(F(<\&_q,W,m)>0)$ и при каждом

выполнения шага 5 либо происходит обновление текущих оптимальных моделей на новые с меньшим значением функции $F(0,*¥\ r_q)$, либо

текущие оптимальные модели остаются неизменными.

Таким образом, предложенный алгоритм позволяет выбрать оптимальные количества смесей для целевой и альтернативной моделей. Кроме того, алгоритм позволяет выбирать начальные приближения.

2.4 Выводы

В главе даны определения целевой и альтернативной моделей распределения векторов признаков. Построена элементарная функция достоверности для вектора признаков, основанная на отношении правдоподобия целевой и альтернативной модели. Также определена дискриминационная величина для каждой пары целевой и альтернативной модели.

Предложен новый метод формирования оценок достоверности распознанного слова, который состоит в использовании значений

элементарной функции достоверности для составляющих векторов признаков и дискриминационных величин.

Сформулированы критерий и метод обучения целевых и альтернативных моделей. Показано, что с помощью метода градиентного спуска можно осуществлять обучение моделей согласно выбранному критерию. Указаны недостатки метода градиентного спуска: эмпирический выбор количества параметров моделей и зависимость от начального приближения. Разработан алгоритм обучения целевых и альтернативных моделей, позволяющий устранить указанные недостатки метода градиентного спуска.

Глава 3. Экспериментальные применения 3.1

Корпус речевых данных FaVoR

Численные эксперименты были выполнены на речевом корпусе данных FaVoR [7], который содержит записи слитной речи 1673 дикторов. Все записи корпуса оцифрованы с частотой дискретизации 22,050 кГц и хранятся в файлах формата Microsoft Wave. Словарь корпуса состоит из 14 слов и содержит цифры от 0 до 9, и служебные слова «да», «нет», «старт» и «стоп». Речевые записи, из которых состоит корпус FaVoR, представляют собой высказывания, состоящие из произнесения персональных идентификационных номеров - четырехзначных чисел, а также образцы произнесения всех слов словаря системы. Корпус FaVoR записан в естественной, достаточно шумной акустико-фоновой обстановке (среднее отношение сигнал/шум равно 15 дБ), с присутствием значительного количества различных незнакомых слов и экстралингвистических событий (кашель, заполненные паузы, смех и т.п.). Наличие заметного количества экстралингвистических событий объясняется, по видимому тем, что значительная часть дикторов не была знакома с компьютерами и процедурами записи речи.

Для каждой речевой записи в корпусе FaVoR имеется аннотация -текстовый файл, содержащий разметки всех произнесенных слов и звуков согласно фонетической (произносительной) транскрипции. Структура разметки отражает временное выравнивание (time-alignment) речевых единиц (слов и звуков) с сигналом. Аннотация состоит из нескольких разделов. Каждый раздел начинается с новой строки, с первой позиции, именем раздела, заключенного в квадратные скобки. В текущей версии использовались следующие разделы: [Text], [Transcription], [Phonemes]. Раздел [Text] содержит правильный литературный текст высказывания (то есть так, как высказывание планировалось диктором, то, что он хотел

сказать, без реальных произносительных ошибок) с заглавными буквами, знаками препинания, сокращениями, цифрами и т.п. Раздел [SRO] содержит нормализованный текст, который соответствует фактическому высказыванию, в котором названия цифры записаны в орфографическом представлении, заглавные буквы заменены на прописные, отсутствуют запятые, точки и другие знаки препинания, которые не соответствуют произношению звуков. Раздел [Phonemes] содержит список фонем, из которых состоит высказывание, с информацией о начале и конце каждой фонемы. Раздел [Words] содержит список произнесенных слов, с информацией о начале и конце каждого слова. Пример аннотации из корпуса FaVoR приведен в Приложении 1.

Используемый набор меток для обозначения звуков русской речи приведен в таблице 5. Мягкость согласного звука обозначается символом апострофа «'», а ударные гласные обозначаются с помощью добавления символа циркумфлекса «^{Л}».

Таблица 5

Система обозначений звуков речи корпуса FaVoR

Метка	Звук	Метка	Звук	Метка	Звук
A	а (безударный)	**V'**	в (мягкий)	p'	п (мягкий)
а^{Л}	а (ударный)	V	в (твердый)	p	п (твердый)
E	е (безударный)	d'	д (мягкий)	**1-'**	р (мягкий)
е^{л}	е (ударный)	D	д (тведый)	г'	р (твердый)
I	и (безударный)	Г	л (мягкий)	**S'**	с (мягкий)
і^{л}	и (ударный)	ш'	м (мягкий)	**S**	с (твердый)
о^{л}	о (ударный)	п'	н (мягкий)	***V***	т (мягкий)
уA	ы (ударный)	п'	н (твердый)	t	т (твердый)
wO	а, е, и, о	**СП'**	ч (мягкий)	sh'	ш (твердый)

В таблице 6 представлены произносительные транскрипции слов базовой системы распознавания речи.

Таблица 6

Произносительная транскрипция слов

Слово	Произносительная транскрипция
Старт	sta^{A}rt
Стоп	sto^{A}p
Да	da^{A}
Нет	n'e^{A}t
Ноль	no^{A}l'
Один	ad'i^{A}n
Два	dva^{A}
Три	tr'i
Четыре	ch'wOtyVe
	ch'ityVe
Пять	p'a^{A}t'
Шесть	she^{A}s't'
Семь	s'e^{A}m'
Восемь	vo^{A}s'im'
	vo^{A}s'wOm'
Девять	d'e^{A}v'wOt'
	d'e^{A}v't'

Корпус данных разделен на 3 части: обучающая выборка, настроечная выборка и тестовая выборка. Обучающая выборка используется для обучения акустических моделей базовой системы распознавания речи. Настроечная выборка предназначена для выбора параметров целевых и альтернативных

моделей функции достоверности. Вероятности правильного распознавания для базовой системы и эффективность предложенных оценок достоверности распознавания производились на тестовой выборке. Характеристики каждой из выборок представлены в таблице 7. В таблице 8 приведены распределения слов по выборкам.

Таблица 7

Характеристики обучающей, настроечной и тестовой выборок

Выборка	Количество дикторов	Количество записей	Количество слов
Обучающая	719	3668	106372
Настроечная	346	2196	63684
Тестовая	325	1464	42456

Таблица 8

Распределения отдельных слов по трем выборкам

Слово	Обучающая выборка	Настроечная выборка	Тестовая
Старт	3668	2196	1464
Стоп	3668	2196	1464
Да	3668	2196	1464
Нет	2668	2196	1464
Ноль	18157	7157	3723
Один	8374	8792	6155
Два	8474	5483	3155
Три	8528	5473	2974
Четыре	8233	5298	3107
Пять	8003	4454	3715
Шесть	8176	4567	3561

Продолжение таблицы 8

Семь	8286	4616	3507
Восемь	7886	4616	3507
Девять	7583	4823	3295

3.2 Базовая система распознавания речи

Базовая система распознавания речи была построена с помощью программного пакета с открытым исходным кодом Hidden Markov Model Toolkit (HTK) фирмы Entropic Reasearch Laboratory [83]. HTK является мощным средством для построения и обработки СММ. Главным образом HTK используется в задачах распознавания речи, распознавания символов, синтеза речи и секвенирования ДНК.

3.2.1 Извлечение векторов признаков речевого сигнала

Входной сигнал преобразуется в последовательность 42-мерных векторов признаков. Каждый вектор признаков состоит из 13 мел-кепстральных коэффициентов, логарифма энергии сигнала, их первых и вторых производных по времени. Параметры алгоритма вычисления векторов признаков установлены следующими:

Коэффициент предварительной коррекции $a = 0.97$.

Длина кадра анализа - 25 мс.

Длина сдвига кадра анализа - 15 мс.

Взвешивание сигнала выполняется с помощью окна Хэмминга

$$w(n) = 0.54 - 0.46 \cos\left(\frac{2\pi n}{N-1}\right), 0 < n < N$$

Количество частотных полос по шкале Мелов - 24. Блок-схема алгоритма вычисления векторов признаков представлена на следующем рисунке.

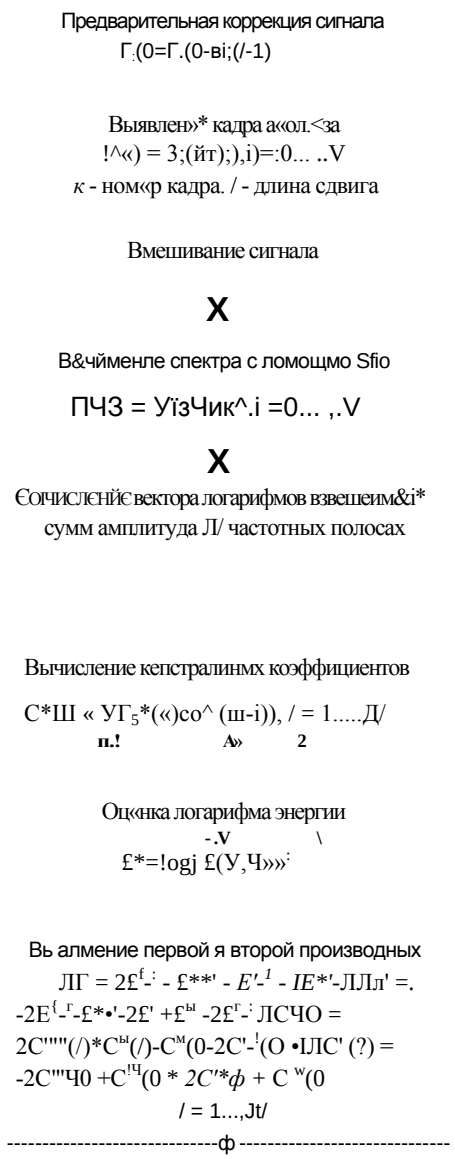


Рисунок 5. Блок-схема алгоритма вычисления векторов признаков речевого сигнала

3.2.2 Акустические модели звуков речи

В качестве базовых единиц для акустического моделирования выбраны контекстно-зависимые реализации фонем - фоны. Контекстная зависимость в данном случае означает влияние предыдущей или следующей фонемы на значения параметров текущей фонемы. Состав фонов для слов словаря системы распознавания определяется их произносительной транскрипцией. Каждый контекстно-зависимый фон имеет один из следующих видов:

- y-x - фон x с учетом предшествующего фона y (левый контекстно-зависимый би фон).
- x+z - фон x с учетом последующего фона z (правый контекстно-зависимый бифон).
- y-x+z - фон x с учетом предшествующего и последующего фонов y И Z соответственно (трифон).

Перечень слов и соответствующих им последовательностей контекстно-зависимых фонов приведен в таблице 9.

Таблица 9

Последовательности контекстно-зависимых фонов для слов

Слово	Последовательность фонов (контекстно-зависимые)
Старт	s+t s-t+a^{A} t-a^{A}+r a^{A}-r+t r-t
Стоп	s+t s-t+o^{A} t-o^{A}+p o^{A}-p
Да	d+a^{A} d-a^{A}
Нет	n'+e^{A} n'-e^{A}+t e^{A}-t
Ноль	n+o^{A} п-o^{A}+Г o^{Л}-Г
Один	a+d' a-d'+i^{A} d'-i^{A}+n i^{A}-n
Два	d+v d-v+a^{A} v-a^{A}
Три	t+r' t-r'+i^{A} r'-i^{A}
Четыре	ch'+i ch'-i+t i-t+y^{A} t-y^{A}+r' y^{A}-r'+e r'-e
	ch'+vvO ch'-wO+t wO-t+y^{A} t-y^{A}+r' y^{A}-r'+e r'-e
Пять	p'+a^{A} p'-a^{A}+t' a^{A}-t'
Шесть	sh+e^{A} sh-e^{A}+s' e^{A}-s'+t' s'-t'
Семь	s'+e^{A} s'-e^{A}+m' e^{A}-m'
Восемь	v+o^{A} v-o^{A}+s' o^{A}-s'+i s'-i+m' i-m'
	v+o^{A} v-o^{A}+s' o^{A}-s'+wO s'-wO+m' wO-m'
Девять	d'+e^{A} d'-e^{A}+v' e^{A}-v'+t' v'-t'
	d'+e^{A} d'-e^{A}+v' e^{A}-v'+wO v'-wO+t' wO-t'

Каждый контекстно-зависимый фон моделируется непрерывным СММ с топологией Бэкиса и структурой из 3 состояний. Функция распределения параметеров в каждом состоянии описывается смесью из 16 нормальных распределений. Для моделирования межсловных пауз используется отдельная СММ, названная модель «Пауза». Модель «Пауза» имеет 3 состояния, функция распределения каждого из которых является смесью 32 нормальных распределений. Топология модели «Пауза» отличается от топологии Бэкиса тем, что допускаются переходы между первым и последним состояниями.

Таким образом, акустическая модель состоит из 58 СММ. Обучение СММ проводилось согласно критерию максимального правдоподобия (1.6) на обучающей выборке с помощью итерационной процедуры Баума-Уелча.

3.2.3 Модель языка для корпуса данных FaVoR

Так как словарь корпуса FaVoR небольшой и все ограничения на допустимые цепочки слов известны, то модель языка была построена с помощью формальной грамматики. Наглядное графическое представление модели языка базового системы распознавания показано на следующем рисунке.

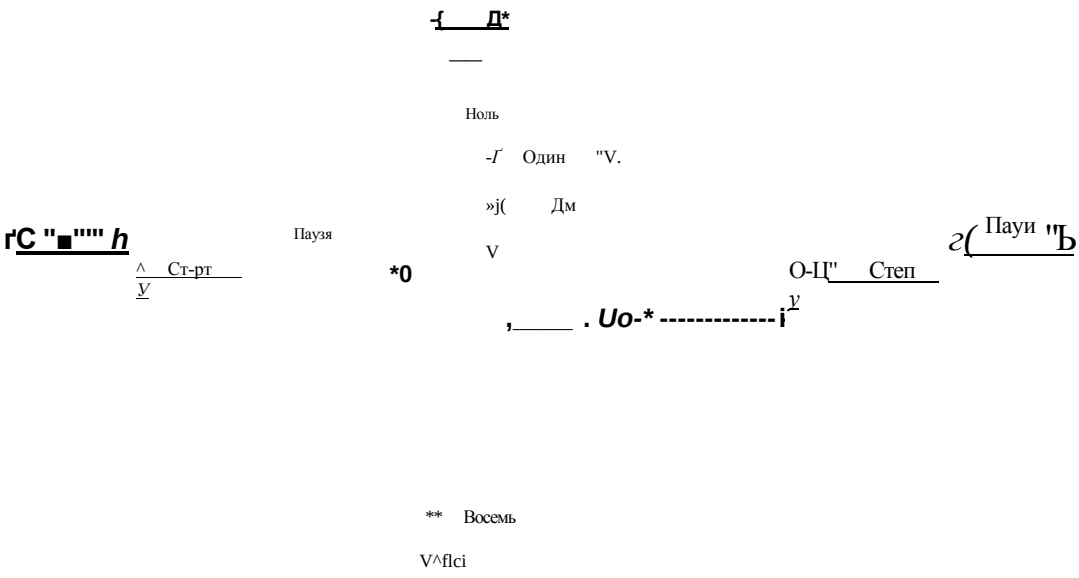


Рисунок 6. Графическое представление модели языка

3.2.4 Эффективность распознавания для базовой системы

Модуль декодирования использует алгоритм Витерби при поиске оптимальной цепочки слов для каждой речевой записи. Для получения оценки эффективности базовой системы распознавания речи каждая распознанная цепочка слов выравнивалась с соответствующей цепочкой слов из файла аннотации методом динамического программирования. Результаты распознавания для слов на настроечной и тестовой выборках приведены в таблице 10. Результаты распознавания для каждого слова словаря системы распознавания речи приведены в Приложении 2.

Таблица 10

Результат распознавания базовой системы

Выборка	Корректные	Пропуски	Замены	Вставки
Настроечная	63149	45	490	1441
Тестовая	40907	34	515	971

Общая пословная ошибка распознавания базовой системы WER (word error rate) на двух выборках

$$\frac{(43 + 34) + \mathbf{(490} + \mathbf{515)} + \mathbf{(1441} + \mathbf{971)}}{63684 + 41456}$$

Количество речевых записей, распознанных без ошибок, составляют 1248 и 836 на настроечной и тестовой выборках соответственно.

3.3 Результаты экспериментов

Так как для слов «старт» и «стоп» отсутствуют ошибки распознавания, то в дальнейшем будем проводить эксперименты только для остальных слов. Соответственно, показатели эффективности предложенных оценок достоверности так же вычисляется без учета слов «старт» и «стоп».

3.3.1 Оценка параметров целевых и альтернативных моделей

Обучение целевых и альтернативных моделей проведено предлагаемым методом обучения (раздел 2.3.3) со значением минимальной суммы компонентов смесей $M_{тш}$=8 и значением £ = 0.01. Для вычисления значения функции ошибки $F(<\&\ ,^{x}¥_q,r_g)$ параметры a_q и b_q установлены равными 1 и 0 соответственно, т. е. функция $R(x,<\&_q,4^{J}_q,T_q)$ имеет вид

В результате обучения всех целевых и альтернативных моделей, установлены следующие факты:

- В подавляющем большинстве случаев, в результате применения алгоритма обучения, число компонентов смесей распределения альтернативной модели больше, чем число компонентов смесей распределения целевой модели. Это факт объясняется тем, что множество / состоит из векторов признаков некорректно распознанных последовательностей векторов признаков, которые были «принудительно» распознаны как слова словаря. Поэтому элементы множества / имеет более широкий разброс по пространству векторов признаков.
- Значения дискриминационных величин для каждого отдельного слова имеют тенденция на убывание на краях. На рис. 7 представлены график зависимости значения дискриминационных величин от состояния для слов «два», «три» и «семь».

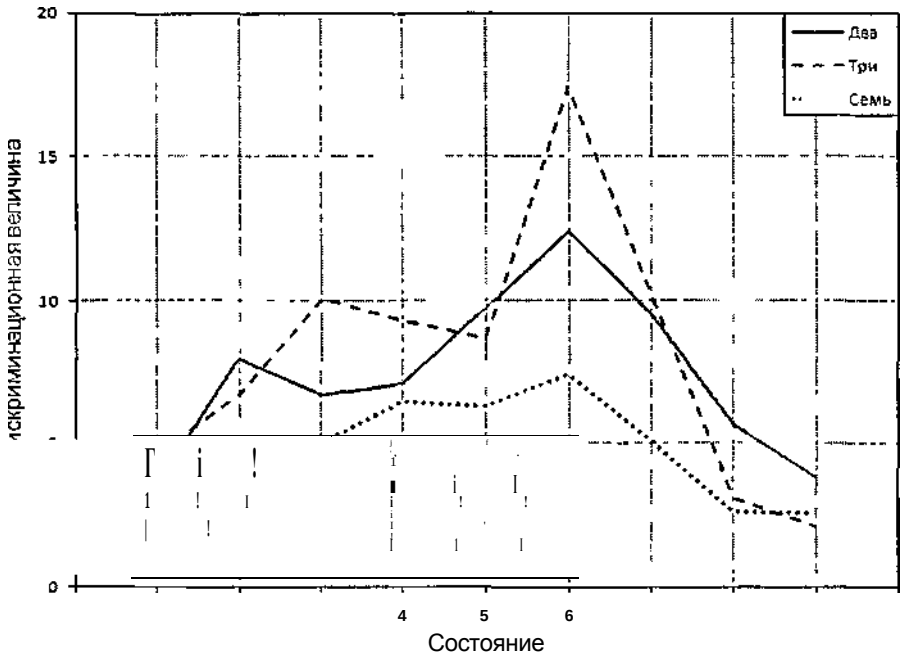


Рисунок 7. График зависимости значения дискриминационной величины от состояния

Первый факт может служить как рекомендация при построении оценок достоверности. Для методов 3-ей группы, т. е. с использованием дополнительных акустических моделей, если необходимо заранее определить число параметров целевых и альтернативных моделей, то целесообразно использовать больше параметров для описания альтернативных моделей, чем для описания целевых моделей.

В качестве примера, результаты обучения целевой и альтернативной моделей для состояния #з СММ контекстно-зависимого фона «n'-e^A+t» на каждом итерации приведены в таблице 11. Числа компонентов смесей распределения целевой и альтернативной моделей обозначаются через M_0 и M, соответственно. Для сравнения проведем обучение целевой и альтернативной моделей состояния $д_3$ СММ «n'-e^A+t» методом градиентного

спуска. Однако алгоритм градиентного спуска требует априори определить числа компонентов смесей распределения целевой и альтернативной моделей и задать их начальные приближения. Поэтому, для каждого выбора числа

компонентов смесей будем проводить обучение 50 раз с различными начальными приближениями. Начальные приближения задаются путем применения алгоритма k-средних и ЕМ-алгоритма на множествах C_q и /

для целевой и альтернативной моделей, соответственно. В таблице 12 приведены результаты применения алгоритма градиентного спуска с разными размерами целевой и альтернативой моделей. Из результатов видно, что полученные значения M_0 и M_x при применении предложенным алгоритмом обучения является оптимальными. Для случаев обучения целевой и альтернативной моделей методом градиентного спуска с большим числом параметров не гарантируются уменьшение функции ошибки *F(0 ^{x}¥$_q$,r).* Кроме того, большое число параметров может привести к

проблеме нехватки данных для обучения.

Таблица 11

Результаты обучения состояния q_3 СММ «n'-e^{A}+t»

предложенным алгоритмом

Итерация	M$_0$	M,	***ПФ$_д$^$_д$,m$_д$)***
0	1	1	0.23882
1	1	2	0.19003
2	1	3	0.16682
3	1	4	0.09316
4	2	4	0.07820
5	2	5	0.06241
6	3	5	0.04352
7	3	6	0.03599
8	3	7	0.02839
9	4	7	0.01676
10	4	8	0.01556

Таблица 12

Результаты обучения состояния $д_3$ СММ «n'-e^A+t»

алгоритмом градиентного спуска

$м_0$	M,	WV'\9)		
		Минимум	Максимум	Среднее
4	4	0.095566	0.156027	0.111625
4	8	0.022771	0.088676	0.050130
6	6	0.028989	0.116700	0.067049
8	8	0.023937	0.095824	0.053019
12	12	0.012459	0.076366	0.031541

3.3.2 Применения предлагаемых методов формирования оценок достоверности

В зависимости от схемы формирования, оценка достоверности для слова может иметь один из следующих видов

$$Cm_A\{W,X) = Yua_tC\{x_t,q_t)$$

$$Cm_G\{W, X) = \exp \pounds\ a_t \ln C\{x_t, q_t)$$

$$Cm_{ю}(Ж,X) = \hat{}Cm_0(u_n,X_u)\ Cm_{AA}(Ф,X) =$$

$$\pm\text{-}\hat{}Cm_A(u_n,X_{un})\ Cm_{GA}\{W,X) =$$

$$\exp\text{fl}\pounds\ln Cm>„,\hat{})$$

$$Cm_{GG}(\hat{},X) = \exp\left(\frac{1}{N}\pounds\ln Cm_G0„,X„)\right)^{>}$$

где $W = u_x...u_N$ - распознанное слово, состоящее из ./V фонов, $X = (x_1,...,x_m)$ ~ последовательность векторов признаков, 6 = (^,,...,(77-) - оптимальная последовательность состояний, $Cm_A(u_n,X_u)$ и $Cm_G(u_n,X_u)$ - оценки достоверности фона u_n

'=•*„

(ᵉ″ *)

V=*. *J*

$X_u = (x_s,...,x_e)$ - соответствующая фону u_n подпоследовательность

векторов признаков, a_t и $JЗ_{n\ t}$ - весовые коэффициенты

$$a=\hat{}A-,l<t<T$$

$$\sum_{/=1}^{2}X$$

$$P_{n,t}=-r\hat{}-,\backslash<n<N,s_n<t<e_n,$$

d - дискриминационная величина состояния q_t, $к>0$ - числовой параметр.

Рассмотрим случаи, когда оценка достоверности для слова вычисляется без учета дискриминационных величин, т.е. при $к = 0$. В таблице 13 представлены показатели частоты ошибок классификации (CER) предложенных методов с выборочным оптимальным значением порога для каждого слова. В таблицах, жирными шрифтами выделены налучшие показатели эффективности. Базовая частота ошибок классификации соответствует случаю, когда все распознанные слова считаются корректно распознанными, и вычисляется по формуле

$$\_ \frac{Ins + Sub}{Num}$$

где *Ins* - количество вставок, *Sub* - количество замен, *Num*— количество распознанных слов.

Таблица 13

Показатели эффективности CER предложенных оценок достоверности

Слово	Частота ошибок классификации CER(%)						
	JJ V-tiJ C/^rn	Cm_A	Cm_c	Cm_{AA}	$^{Cf}n_{AG}$	^{Cm}GA	Cm_{GG}
Да	13.290	9.058	7.330	7.449	7.688	7.569	**6.734**
Нет	13.111	4.589	3.695	3.695	4.052	3.159	**2.980**
Один	2.583	2.343	2.104	1.849	2.296	**1.769**	1.801
Два	2.620	2.005	**1.714**	1.973	2.038	2.329	1.876
Три	5.458	3.834	4.808	**3.671**	4.386	3.899	4.711
Четыре	0.804	0.804	0.772	**0.611**	0.772	0.675	0.708
Пять	2.990	2.858	2.964	2.673	**2.461**	2.646	2.964
Шесть	2.660	2.550	2.660	**2.194**	2.660	2.660	2.660
Семь	6.065	4.879	5.795	4.528	5.364	**4.097**	4.987
Восемь	0.148	0.118	0.118	**0.089**	0.118	0.148	0.118
Девять	1.210	1.179	1.117	1.210	**1.055**	1.210	1.148
Ноль	3.350	3.088	3.245	2.931	3.219	**2.643**	3.219
Всего	3.672	2.775	2.792	2.469	2.750	**2.466**	2.617

Результаты подтверждают эффективность предложенных оценок достоверности. Кроме того, общая частота ошибок классификации получается меньше при использовании двухуровневой нормализации. Это объясняется чем, что эти варианты позволяют подавлять «скачки» составляющих значений элементарных функции достоверности внутри

каждого фона. Кривые характеристики ROC для слов «ноль», «один» и «семь» метода $Cm_{GA}(W,X)$ представлены на следующем рисунке.

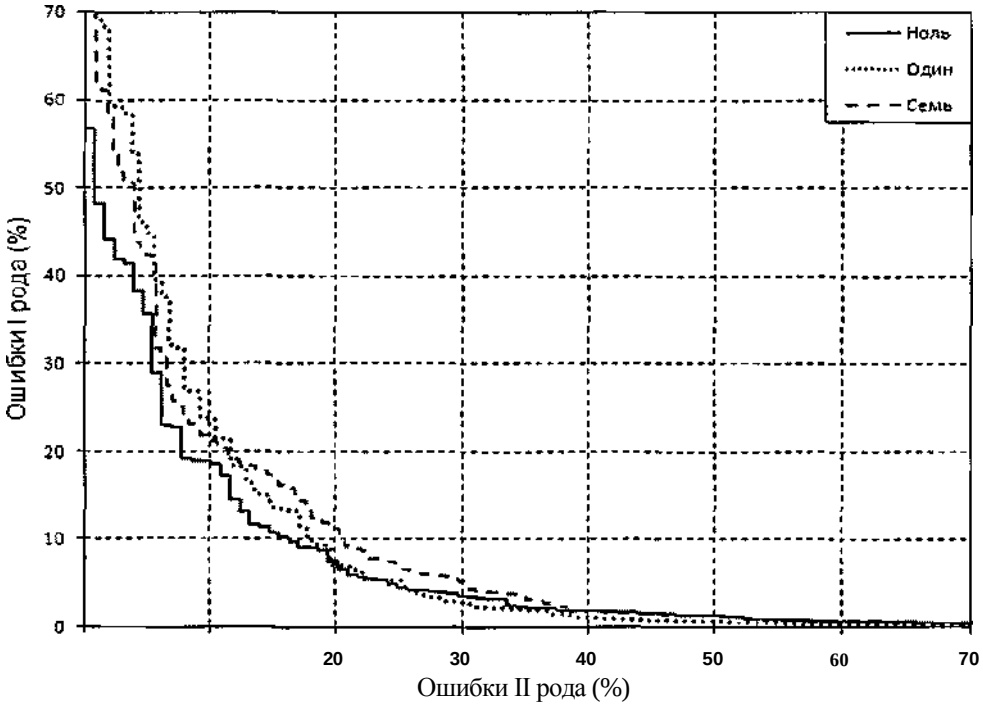


Рисунок 8. Кривые характеристики оценки достоверности $Cm_{CA}(W,X)$ со значением $к$ = Одля слов «ноль», «один» и «семь»

Однако выбор оптимального значения порога для каждого отдельного слова словаря системы распознавания речи представляется не оптимальным с точки зрения практического применения. Например, потому, что при добавлении нового слова, которое состоит из уже имеющихся в системе смоделированных контекстно-зависимых фонов, придется проводить эксперименты по нахождению оптимального значения порога для этого слова. Таким образом, в дальнейшем будем оценить эффективность предложенных оценок достоверности с использованием единого значения порога для всех слов. На рис. 9 и 10 представлены кривые характеристики (ROC) оценок достоверности при А- = 0. В таблице 14 приведены показатели EER и CER предложенных оценок достоверности при $к$ = 0.

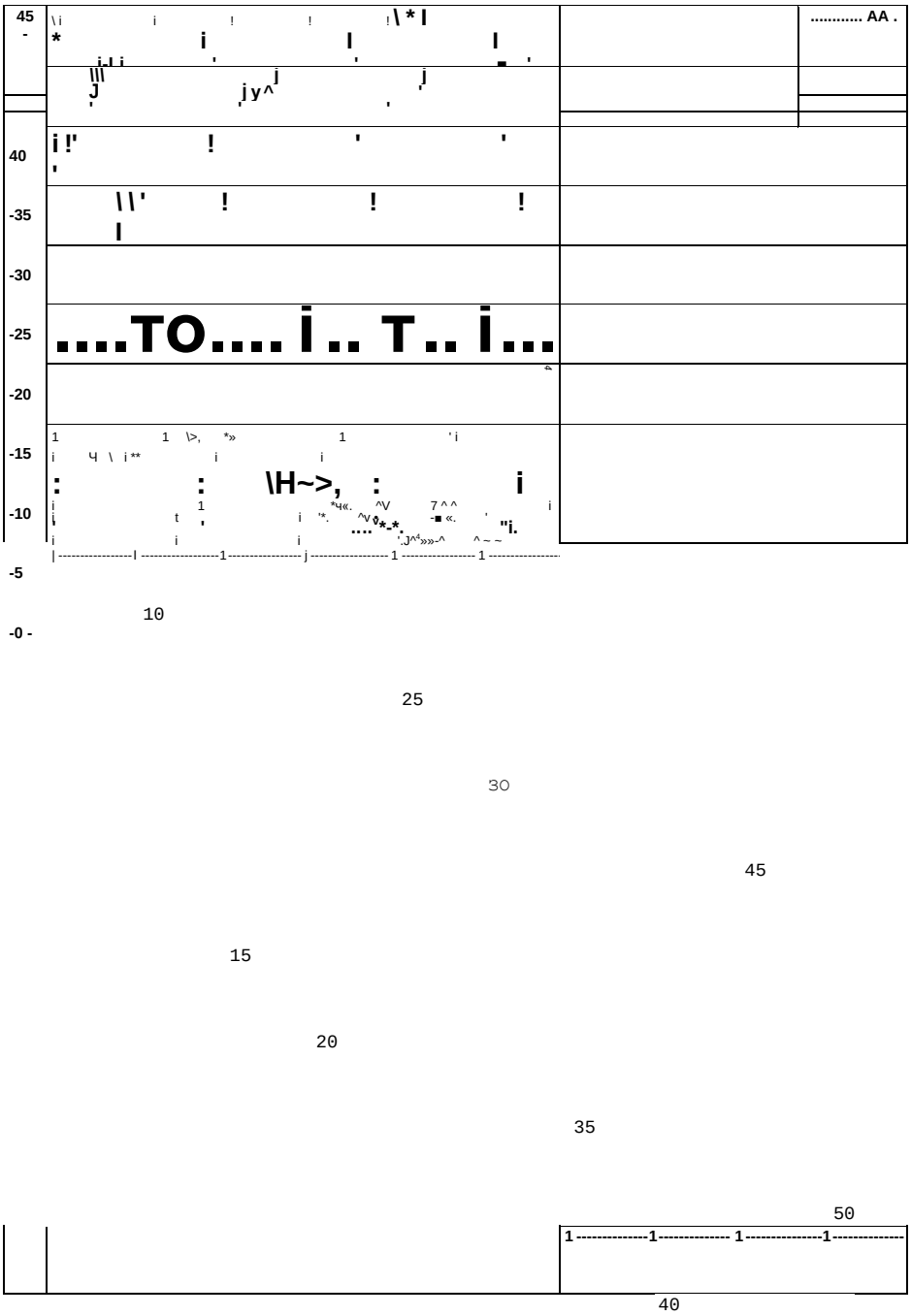


Рисунок 9. Кривые характеристики оценок достоверности $Cm_A(W,X)$, $Cm_{AA}(W,X)$ и $Cm_{AG}(W,X)$ со значением $k = 0$

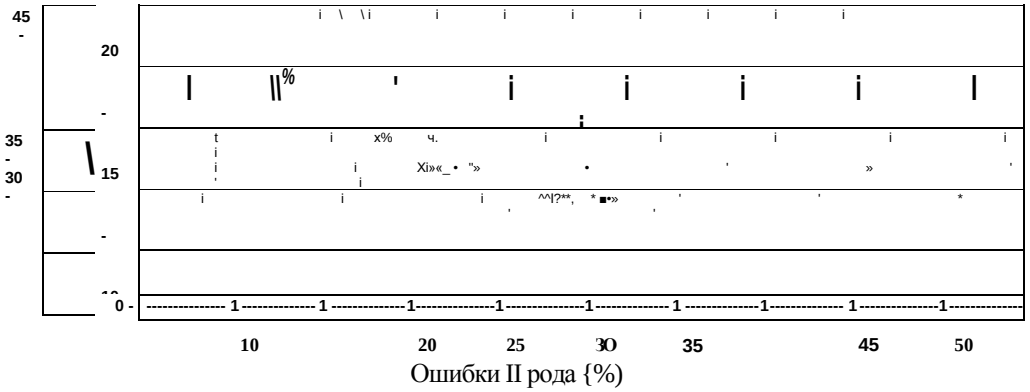


Рисунок 10. Кривые характеристики оценок достоверности $Cm_G(W,X)$, $Cm_{GA}(W,X)$ и $Cm_{GG}(W,X)$ со значением $k = 0$

Таблица 14

Показатели эффективности предложенных оценок достоверности при $\kappa = 0$.

Оценка достоверности	ERR(%)	CER(%)
$Cm_A(W,X)$	14.469	3.077
$Cm_u(W,X)$	12.314	**2.746**
$Cm_{AG}(W,X)$	13.392	3.114
$Cm_G(W,X)$	13.392	3.166
$Cm_{GA}(W,X)$	12.380	2.792
$Cm_{GG}(W,X)$	**12.045**	2.916

Как и в случае выбора значения порогов, предполагается целесообразно использовать одно и то же значение κ при вычислении весовых коэффициентов для всех слов. В таблице 15 приведены показатели эффективности (ERR и CER) оценок достоверности $Cm_{AA}(W,X)$ и $Cm_{GG}(W, X)$. Из этих результатов можно сделать вывод о том, что использование весовых коэффициентов с учетом дискриминационных величин приводит к улучшению показателей эффективности.

Таблица 15

Показатели эффективности оценок достоверности $Cm_{AA}(W,X)$ и $Cm_{GG}(W,X)$ при £ = 0 и $\kappa = 1$

Оценка достоверности	κ	EER(%)	CER(%)
$Cm_{AA}\{W,X)$	0	12.314	2.746
$Cm_{AA}i\}V,X)$	1	12.045	**2.627**
$Cm_{GG}(W,X)$	0	12.045	2.916
$Cm_{GG}(lV,X)$	1	**11.723**	2.901

В результате экспериментов установлено, что минимальное значение показатели EER= 11.508% достигается при использовании оценки достоверности $Cm_{GG}(W,X)$ со значением $к$ =0.7. В таблице 16 представлены результаты для всех предложенных оценок достоверности. На рис. 11 показаны графики зависимости значения показатели EER от значения $к$ для оценок достоверности $Cm_{GG}(W,X)$ и $Cm_{AA}(W,X)$.

Таблица 16

Минимальные значения EER предложенных оценок достоверности

Оценка достоверности	$к$	EER(%)
$Cm_A(W,X)$	1.6	13.258
$Cm_{AA}(W,X)$	0.5	11.911
$Cm_{AG}(W,X)$	1.5	12.723
$Cm_G(W,X)$	1.6	12.314
$Cm_{GA}(W,X)$	0.8	11.911
$Cm_{GG}(W,X)$	0.7	11.508

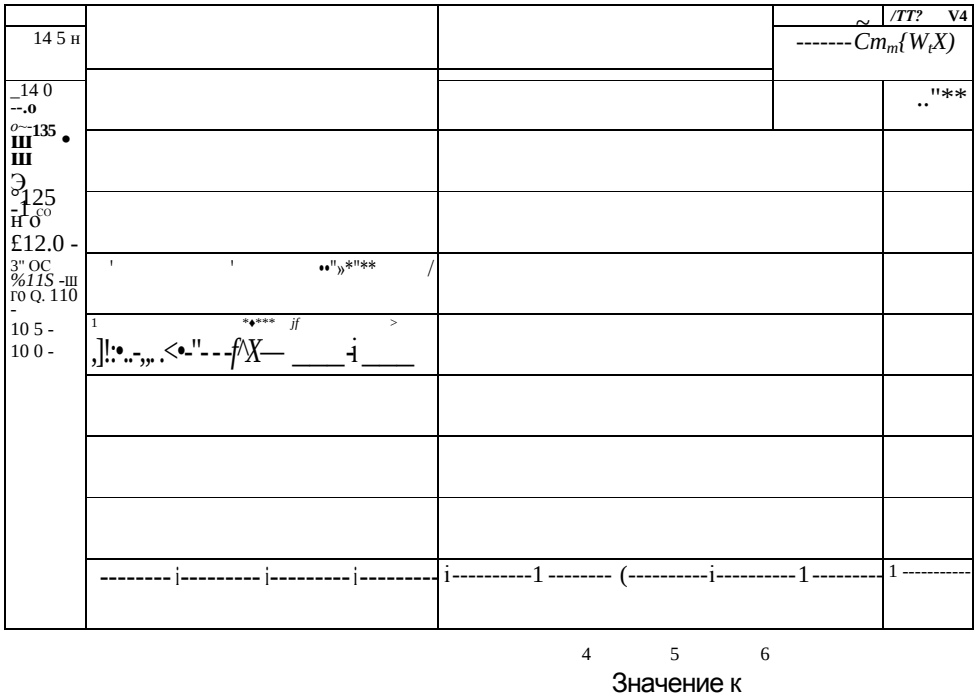


Рисунок 11 График зависимости значения EER оценок достоверности $Cm_{AA}(W,X)$ и $Cm_{GG}\{W,X)$ от значения $к$

Аналогичные эксперименты проведены для исследования зависимости показателя CER предложенных оценок достоверности от значения $к$. Наилучшее значение показателя CER=2.533% достигается при использовании оценки достоверности $Cm_{AA}(W,X)$ с параметром $к$ = 6.7.

Минимальные значения CER для каждой из оценок достоверности представлены в таблице 17. Графики зависимости значения показатели CER от значения $к$ для оценок достоверности $Cm_{AA}(W,X)$ и $Cm_{A}(W,X)$ показаны нарис. 12.

Таблица 17

Минимальные значения CER предложенных оценок достоверности

Оценка достоверности	*к*	CER(%)
$Cm_A(W,X)$	2.7	2.654
$Cm_{AA}(W,X)$	6.7	2.533
$Cm_{AG}(W,X)$	9.2	2.822
$Cm_G\{W,X)$	7.6	2.849
$Cm_{GA}(W,X)$	3.9	2.711
$Cm_{GG}(W,X)$	0.5	2.896

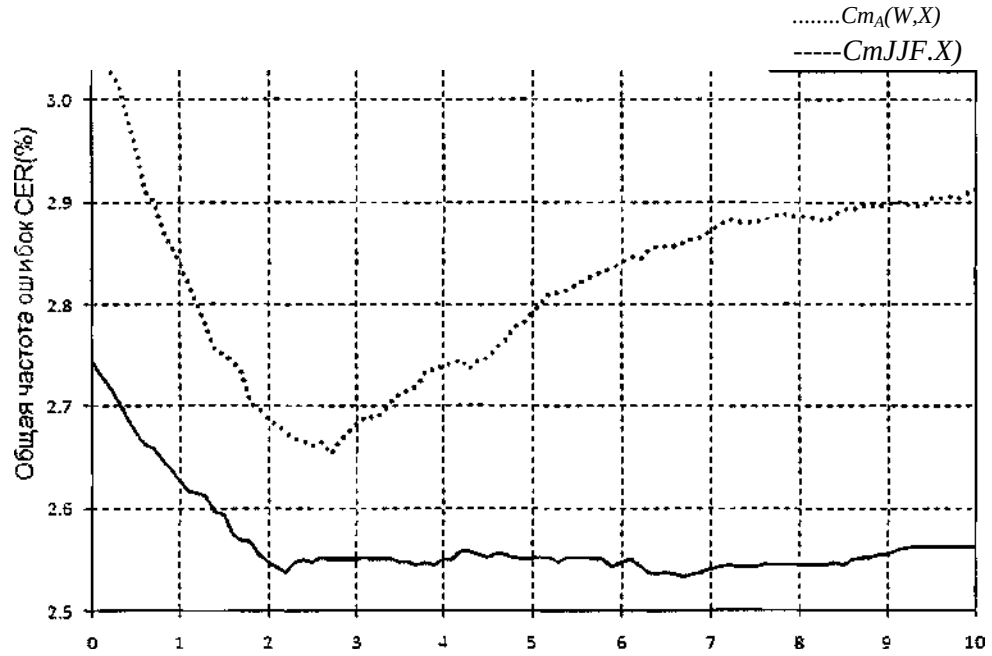


Рисунок 12. График зависимости значения EER оценок достоверности $Cm_{AA}(W,X)$ и $Cm_{GG}(W,X)$ от значения *к*

3.3.3 Сравнение эффективности предложенного метода с известными оценками достоверности

Проведем сравнения эффективности предложенного метода формирования оценок достоверности с существующими методами. В

качестве метода первой группы возьмем нормированную акустическую оценку

$$Cm_{NAS}\{W,X) = jP\{X\backslash^{\wedge}) ,$$

где *T*- длина последовательности векторов признаковI, /^ - СММ слова *W*. Показатели эффективности EER и CER при применении нормированной акустической оценки на тестовой выборке составляют 40.595%) и 3.640%> соответственно. На рис. 13 представлено наглядное сравнение оценки $Cm_{NAS}(W,X)$ с предложенными оценками $Cm_{AA}\{W,X)$ *(к = 6.1)* и $Cm_{GG}(W,X)$ *(к =* 0.7)).

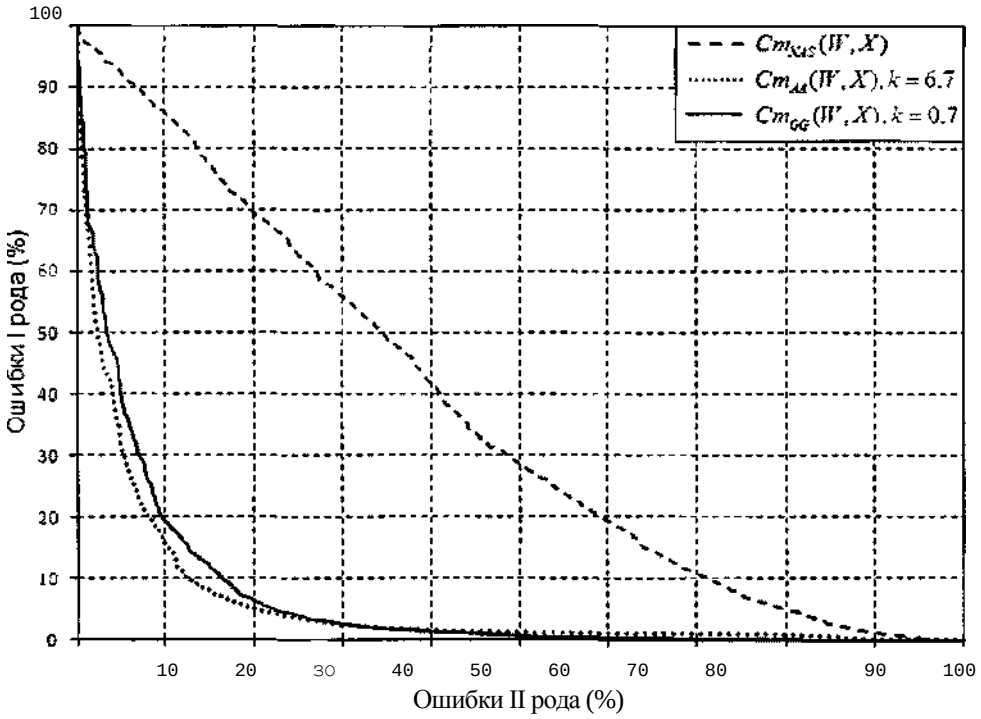


Рисунок 13. Кривые характеристики ROC оценок достоверности $Cm_{NAS}(W,X)$, $Cm_{AA}(W,X)$ *{к* = 6.7) и $Cm_{GG}(W,X)$ (А-= 0.7)

Для методов второй группы, рассмотрим оценку достоверности $Cm_{MAX}(W,X)$, описанную в разделе 1.2.2,

$$Cm_{MAX}(W,X) = \max_{1} \quad \% P(W,s',e'\backslash X)$$
$$[w,s,e\hat{}$$
$$s'<t<e'$$

Результаты применения оценки достоверности $Cm_{MAX}(W,X)$ на корпусах данных Nab20k и Nab64k [71] представлены в таблице 18. На каждом из этих корпусах получены 18.6% и 21.7% относительного улучшения по сравнению с базового показателя $Base_{CER}$. В то время, предложенная оценка достоверности $Cm_{AA}(W,X)$ при $к$ = 6.7 дает 31.0% относительного улучшения показателя $Base_{CER}$ на тестовой выборке.

Таблица 18

Сравнение эффективности оценок достоверности

$Cm_{MAX}(W,X)$ и $Cm_{AA}\{W,X)$

Оценка достоверности	Корпус данных	$Base_{CER}$ (%)	CER (%)	Относительное улучшение(%)
$Cm_{MAX}(JV,X)$	Nab20k	11.3	9.2	18.6
$Cm_{MAX}\{W,X)$	Nab64k	9.2	7.2	21.7
$Cm_{AA}(JV,X)(k = 6.7)$	FaVoR	3.7	2.5	31.0

В [15] предложен метод формирования оценок достоверности, относящийся к третьей группе. Для этого построены целевая и альтернативная модели каждого фона $и$. Оценка достоверности для слова вычисляется согласно формуле

$$Cm_{LR}(lV,X) = Z_{n=i}^{N} \frac{\tau \quad \mathbf{i},}{nxjxj\hat{}} \quad a_{„} —\log \text{------} \frac{''\quad\quad''}{\backslash^{n}}$$

где N - количество составляющих фонов слова W, X_u - соответствующая фону u_n подпоследовательность векторов признаков, T_u - длина X_u, $Я_u$ и

$\bar{Л}_u$ - целевая и альтернативная модели фона u_n, $a_n > 0 (£ a_n = 1)$ - весовые коэффициенты. В результате экспериментов [15] получено 9% относительного улучшения показателя CER.

Таким образом, можно сделать вывод о том, что предложенный метод в работе формирования оценки достоверности обладает более высокой эффективностью по сравнению с существующими методами.

3.4 Выводы

В главе дано описание речевого корпуса данных FaVoR, на котором проводилось численное исследование эффективности предложенных оценок достоверности. Приведены описания составляющих модулей базовой системы распознавания речи и значение эффективности для базовой системы распознавания речи.

Проведены эксперименты для сравнения эффективности предложенного алгоритма обучения целевых и альтернативных моделей с алгоритмом градиентного спуска. Результаты показали, что предложенный алгоритм обучения обеспечивает лучший выбор количества смесей целевых и альтернативных моделей.

При анализе результатов обучения целевых и альтернативных моделей установлено, что значения дискриминационных величин для каждого отдельного слова имеют тенденцию к убыванию на краях реализации слова, что подтверждает известный эмпирический подход к взвешиванию оценок правдоподобия данных для отдельных кадров анализа.

Проведены исследования эффективности предложенных методов формирования оценок достоверности. Получены показатель равной частоты ошибок EER—11.508%, а относительное улучшение базового показателя частоты ошибок классификации составляло 31%. Установлено, что

двухуровневые методы формирования оценок достоверности превосходят одноуровневые методы.

Показано, что по сравнению с известными оценками достоверности, предложенные оценки достоверности обладают более высокой эффективностью.

Заключение

В диссертационной работе представлены результаты исследований и разработки алгоритмов построения оценок достоверности для систем распознавания речи.

Основные результаты диссертационной работы заключаются в следующем:

1. Проведено исследование современных методов построения систем распознавания на основе вероятностного подхода.
2. Проведен анализ существующих подходов к формированию оценок достоверности для систем распознавания речи.
3. Введены определения целевых и альтернативных моделей распределения векторов признаков речевого сигнала и приведен способ построения элементарной функции достоверности для вектора признаков. Предложен новый метод построения оценок достоверности для систем распознавания речи, который основан на использовании значений элементарной функции от составляющих векторов признаков.
4. На основе предложенного метода построения элементарной функции достоверности для вектора признаков разработан алгоритм построения целевых и альтернативных моделей, который позволяет решить проблему выбора количества параметров этих моделей.
5. Выполнена практическая реализация и проведены численные измерения показателей эффективности предложенных методов и алгоритмов. Результаты экспериментов показали более высокую эффективность предложных в работе оценок достоверности по сравнению с известными оценками.

Приложение №1

Пример аннотации из корпуса FaVoR.

[TEXT]

СТАРТ 00145 00123 шесть два четыре пять девять три один семь ноль восемь ДА НЕТ 00145 СТОП

[SRO]

старт ноль ноль один четыре пять ноль ноль один два три шесть два четыре пять девять три один семь ноль восемь да нет ноль ноль один четыре пять стоп

[Phonemes]

128 29312 sil
29312 30848 s
30848 32896 t
32896 34944 а^Л
34944 36480 Г
36480 37248 t
37248 56704 sil
56704 58752 n
58752 60800 0^А
60800 63616 Г
63616 65152 n
65152 67456 0^А
67456 72576 Г
72576 74368 a
74368 76672 d'
76672 78464 i^А

78464 81024 n

81024 83072 ch'

83072 84608 w0

84608 85376 t

85376 87168 y^{A}

87168 89216 ɾ'

89216 92288 e

92288 93312 p'

93312 95872 a^{A}

95872 100736 t'

100736 116864 sil

116864 118400 n

118400 121984o^{A}

121984 125056 ɾ

125056 126592 n

126592 129408o^{A}

129408 1312001'

131200 134272 a

134272 136576 d'

136576 138624 i^{A}

138624 142464 n

142464 143232 d

143232 145792v

145792 150912 a^{A}

150912 1519361

151936 153984 ɾ'

153984 158080 i^{ɹ}

158080 189824 sil

189824 193152 sh

193152196480 eʲ

196480 198784 s*

198784 204160 t'

204160 207744 sil

207744 209536 d

209536 212352v

212352 217728 aʲ

217728 229504 sil

229504 232064 ch'

232064 234112 i

234112 234880 t

234880 238208 yʲ

238208 240000 ɾ'

240000 244864 ɛ

244864 256384 sil

256384 257664 p'

257664 261504 aʲ

261504 267136 ɾ'

267136 282496 sil

282496 284032 d'

284032 287360 eʲ

287360 289152 V

289152 295296 f

295296 306560 sil

306560 3075841

307584 309120 ɾ*

309120 314752 iʲ

314752 330880 sil

330880 333184 a

333184 334976 d*

334976 337536 iᴬ

337536 341376 n

341376 356480 sil

356480 360320 s'

360320 363904 eᴶᴵ

363904 369024 m'

369024 383872 sil

383872 384896 n

384896 388480 oᴬ

388480 3910401'

391040 409984 sil

409984 411776v

411776 413568 oᴶᴵ

413568 416896 s'

416896 417920 wO

417920 419712 m'

419712 434560 sil

434560 436096 d

436096 442240 aᴬ

442240 470912 sil

470912 471936 n'

471936 474496 eᴬ

474496 4760321

476032 500352 sil

500352 501888n

501888 504448 oᴬ

504448 505728 Г

505728 507520 n

507520 510080оᴬ

510080 5113601'

5113605131522а

513152 515200(1'

515200 516736 іᴶᴵ

516736 518272 п

518272 520832 ch*

520832 522368 і

522368 523392 t

523392 525952 уᴶᴵ

525952 527744 г'

527744 530816 є

530816 531840 р'

531840 534400 аᴶᴵ

534400 538496 f

538496 559744 sil

559744 561536 s

561536 563328 t

563328 566656 оᴶᴵ

566656 568704 p

568704 573439 sil

[Words]

128 29312-

29312 37248старт

37248 56704-

56704 бЗбІбноль

63616 72576ноль

72576 81024один

81024 92288четыре

92288 100736 пять

100736116864

116864125056 ноль

125056131200 ноль

131200142464 один

142464150912 два

150912158080 три

158080189824

189824204160 шесть

204160207744

207744217728 два

217728229504

229504244864 четыре

244864256384

256384267136 пять

267136282496

282496295296 девять

295296306560

306560314752 три

314752330880

330880341376 один

341376356480

356480369024 семь

369024383872

383872391040 ноль

391040409984

409984419712 восемь

419712434560

434560442240	да
442240470912	
470912476032	нет
476032500352	
500352505728	ноль
505728511360	ноль
511360518272	один
518272530816	четыре
530816538496	пять
538496559744	
559744568704	стоп
568704-1	

Приложение №2

Таблица

Результат распознавания на настроечной выборке

Слово	Корректные	Замены	Вставки	Пропуски
Старт	2196	0	0	0
Стоп	2196	0	0	0
Да	2176	106	209	1
Нет	2188	18	118	0
Ноль	7115	20	162	9
Один	8735	52	143	5
Два	5348	21	206	3
Три	5409	48	210	5
Четыре	5274	26	13	2
Пять	4411	54	133	3
Шесть	4558	20	54	0
Семь	4586	79	127	2
Восемь	4209	7	29	1
Девять	4748	39	37	14

Таблица 20

Результат распознавания на тестовой выборке

Слово	Корректные	Замены	Вставки	Пропуски
Старт	1464	0	0	0
Стоп	1464	0	0	0
Да	1455	107	116	0
Нет	1458	22	198	0
Ноль	3693	21	107	4
Один	6110	75	87	6
Два	ЗОН	11	70	11
Три	2910	48	120	1
Четыре	3084	19	6	0
Пять	3666	37	76	6
Шесть	3550	30	67	2
Семь	3485	115	ПО	1
Восемь	3372	2	3	0
Девять	3185	28	11	3

Библиография

1. Нгуен М. Т. Оценка достоверности результатов автоматического распознавания речи // Труды Института системного анализа РАН. Динамика неоднородных систем, 2006, в. 10(2), с. 405-414
2. Нгуен М. Т. Обнаружение новых слов и невербальных событий при распознавании речи // Модели, методы, алгоритмы и архитектуры систем распознавания речи, 2006, с. 119-137
3. Нгуен М. Т. Построение оценок достоверности результатов распознавания речи с использованием альтернативных моделей // Сборник докладов 13-ой Всероссийской конференции «Математические методы распознавания образов», 2007, с. 370-371
4. Нгуен М. Т., Чучупал В. Я. Верификация результатов автоматического распознавания речи // Сборник трудов XIX сессии Российского Акустического Общества, 2007, Т. 3. с. 63-67
5. Nguyen M. T., Chuchupal V. J. Word verification method for automatic speech recognition *II* Proceedings of the XII International Conference "Speech and Computer" Specom'2007, 2007, V. 1, p. 152-156
6. Nguyen M. T., Chuchupal V. J. Word confidence measure based on frame likelihood score *II* Pattern recognition and image analysis. Advances in mathematical theory and application, 2008, N. 3, p. 431-433
7. Десятчиков А. А., Ковков Д. В., Лобанцов В. В., Маковкин К. А., Матвеев И. А., Мурынин А. Б., Чучупал В. Я. Комплекс Алгоритмов Для Устойчивого Распознавания Человека // Известия РАН, Теория и Системы Управления, 2006, с. 119-130
8. Обжелян Н. К., Трунин-Донской В.Н Машины, которые говорят и слушают // Кишинев, Штиница, 1987

9. Aho A. V., Ullman J. D. The Theoiy of Parsing, Translation and Computing *II* Prentice Hall, 1972
10. Atal B. S., Schroeder M. R. Predictive Coding of Speech Signal *II* Proceedings of the International Congress on Acoustic, 1968
11. Bahl L. R., Jelinek F., Mercer R. L. A Maximum Likelihood Approach to Continuous Speech Recognition *II* IEEE Transactions on Pattern Analysis and Machine Intelligence, 1983, pp. 179-190
12. Baum L. E. An Inequality and Associated Maximization Technique in Statistical Estimation of Probabilistic Functions of a Markov Process *II* Inequalities, 1972, V. 3, pp. 1-8
13. Benitez M. C, Rubio A., Toïte A. Different Confidence Measures for Word Verification in Speech Recognition *II* Speech Communication, 2000, V. 32, pp. 79-94
14. Bilmes J. A. A Gentle Tutorial of the EM Algorithm and Its Application to Parameter Estimation for Gaussian Mixture and Hidden Markov Models, 1998
15. Bouwman G., Boves L., Koolwaaij J. Weighting Phone Confidence Measure for Automatic Speech Recognition *II* Workshop on Voice Operated Telecom Services, 2000, pp. 59-62
16. Charlet D. Optimizing Confidence Measure Based on HMM Acoustical Rescoring *II* Proceedings of the ISCA Tutorial and Research Workshop ARS2000, 2000, pp. 203-206
17. Chase L. Word and Acoustic Confidence Annotation for Large Vocabulary Speech Recognition *II* Proceedings of the European Conference on Speech Communication and Technology, 1997, pp. 815-818
18. Cox S., Rose R. Confidence Measures for the Switch-board Database *II* Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, 1996, pp. 511-514

19. Davis K. H., Biddulph R., Balashek S. Automatic Recognition of Spoken Digits *II* The Journal of the Acoustical Society of America, 1952, V. 24,1. 6, pp. 637-642
20. Demuynck K., Van Compemolle D., Wambacq P. Doing Away with the Viterbi Approximation *II* Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, 2002, pp. 717-720
21. Deriven M. Dynamic Bayesian Networks for Speech Recognition *II* Proceedings of the National Conference on Artificial Intelligence, 2002, pp. 981-981
22. Egan J. P. Signal Detection Theory and ROC Analysis *II* Academic Press, 1975
23. Eide E., Gish H., Jeanrenaud P., Mielke A. Understanding and Improving Speech Recognition Performance Through the Use of Diagnostic Tools *II* Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, 1995, pp. 221-224
24. Erzin E., Cetin A. E., Yardfmci Y. Subband Analysis for Robust Speech Recognition in the Presence of Car Noise *II* Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, 1995, pp. 417-420
25. Fabian T., Lieb R, Gunther R., Matthias T. Impact of Word Graph Density on the Quality of Posterior Probability Based Confidence Measures *II* Proceedings of the European Conference on Speech Communication and Technology, 2003, pp. 917-920
26. Fawcett T. An Introduction to Roc Analysis *II* Pattern Recognition Letters, 2006, pp. 861-874
27. Franzini M., Witbrock M., Lee K. A Connectionist Approach to Continuous Speech Recognition *II* Proceedings of the International Conference on Acoustics, Speech and Signal Processing, 1989, pp. 425-428

28. Furui S. Fifty Years of Progress in Speech and Speaker Recognition *II* The Journal of the Acoustical Society of America, 2004, V. 116, I. 4, pp. 2497-2498
29. Gold B., Morgan N. Speech and Audio Signal Processing *II* John Wiley and Sons, 2000
30. Gowdy J. N., Tufekci Z. Mel-scaled Discrete Wavelet Coefficients for Speech Recognition *II* Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, 2000, pp. 1351-1354
31. Harrison T., Fallside F. A Connectionist Model for Phoneme Recognition in Continuous Speech *II* Proceedings of the International Conference on Acoustics, Speech and Signal Processing , 1989, pp. 417-420
32. Huang X. D., Ariki Y., Jack M. A. Hidden Markov Models for Speech Recognition *II* Edinburgh University Press, 1990
33. Humphrys M. Introduction to Artificial Intelligence, 2008, http://www.computing.dcu.ie/~humphrys/ca300/index.html
34. Hunt A., McGlashan S. Speech Recognition Grammar Specification Version 1.0//W3C, 2004
35. Itakura F., Saito S. Analysis Synthesis Telephony Based on the Maximum Likelihood Method *II* Proceedings of the International Congress on Acoustic, 1968, pp. 17-20
36. Jelinek F. Statistical Method for Speech Recognition *II* MIT Press, 1997
37. Jelinek F. The Development of an Experimental Discrete Dictation Recognizer *II* Proceedings of the IEEE, 1985, pp. 1616-1624
38. Jia B., Zhu X., Luo Y., Hu D. Utterance Verification Using Modified Segmental Probability Model *II* Proceedings of the European Conference on Speech Communication and Technology, 1999, pp. 45-48
39. Jiang L., Huang X. D. Vocabulary-independent Word Confidence Measure Using Subword Features *II* Proceedings of the International Conference on Spoken Language Processing, 1998

40. Jurafsky D., Martin J. H. Speech and Language Processing *II* Prentice Hall, 2008
41. Kemp T., Schaaf T. Estimating Confidence Using Word Lattices *II* Proceedings of the European Conference on Speech Communication and Technology, 1997, pp. 827-830
42. Kim K., Youn D. H., Lee C Evaluation of Wavelet Filters for Speech Recognition *II* Proceedings of the IEEE International Conference on Systems, Man and Cybernetics, 2000, v. 4, pp. 2891-2894
43. Levinson S. E. Continuously Variable Duration Hidden Markov Models for Automatic Speech Recognition *II* Computer Speech and Language, 1986, pp. 29-45
44. Lleida E., Rose R. C Efficient Decoding and Training Procedure for Utterance Verification in Continuous Speech Recognition *II* Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, 1996, pp. 507-510
45. Lleida E., Rose R. C Utterance Verification in Continuous Speech Recognition : Decoding and Training Procedures *II* IEEE Transactions on Speech and Audio Processing, 2000, pp. 126-139
46. Macherey K., Bender 0., Ney H. Multi-level Error Handling for Tree-Based Dialogue Course Management *II* Proceedings of the ISCA Tutorial and Research Workshop on Error Handling in Spoken Dialogue Systems , 2003, pp. 123-128
47. Markel J. D., Gray A. H. Linear Prediction of Speech *II* Springer-Verlag, 1976, pp. 31-35
48. Martin A., Doddington G., Kamm T., Ordowski M., Pryzybocki M. The DET Curve in Assessment of Detection Task Performance *II* Proceedings of the European Conference on Speech Communication and Technology, 1997, pp. 1895-1898

49. Mathan L., Miclet L. Rejection of Extraneous Input in Speech Recognition Applications, Using Multi-layer Perceptrons and the Trace of HMMs *II* Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, 1991, pp. 93-96
50. Moreau N., Jouvet D. Use of a Confidence Measure Based on Frame Level Likelihood Ratios for the Rejection of Incorrect Data *II* Proceedings of the European Conference on Speech Communication and Technology, 1999, pp. 291-294
51. Neti C V., Roukos S., Eide E. Word-based Confidence Measures as a Guide for Stack Search in Speech Recognition *II* Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, 1997, pp. 883-886
52. Ney H., Martin S., Wessel F. Statistical Language Modeling Using Leaving-one-out *II* Corpus-based Methods in Language and Speech Processing, 1997, pp. 174-207
53. Normadin T., Lacouture R., Cardin R. MMIE Training for Large Vocabulary Continuous Speech Recognition *If* Proceedings of the International Conference on Acoustics, Speech and Signal Processing , 1994, pp. 1367-1370
54. Picone J. W. Signal Modeling Teclmiques in Speech Recognition *II* Proceedings of the IEEE, 1993, pp. 1215-1247
55. Pinto J., Sitaram R. N. V. Confidence Measures in Speech Recognition Based on Probability Distribution of Likelihoods *II* Proceedings of the European Conference on Speech Communication and Technology Interspeech'2005, 2005, pp. 3001-3004
56. Rabiner L. R. A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition *II* Proceedings of the IEEE, 1989, pp. 257-286

57. Rabiner L. R.$_5$ Juang B. H. Fundamentals of Speech Recognition *II* Prentice Hall, 1993
58. Rabiner L. R., Juang B. H., Levinson S. E., Sondhi M. M Recognition of Isolated Digits Using Hidden Markov Models with Continuous Mixture Densities *II* AT&T Technical Journal, 1985, pp. 1211-1234
59. Rahim M. G., Lee C. H. Discriminative Utterance Verification for Connected Digits Recognition *II* IEEE Transactions on Speech and Audion Processing, 1997, pp. 266-277
60. Razik J., Mella 0., Fohr D., Haton J. P. Local Word Confidence Measure Using Word Graph and N-Best List *II* Proceedings of the European Conference on Speech Communication and Technology, 2005, pp. 3369-3372
61. Robinson A. J., Fallside F. A Dynamic Connectionist Model for Phoneme Recognition *II* Neural Networks from Models to Applications, 1988, pp. 541-550
62. Rose R. C, Juang B. H., Lee C. H. A Training Procedure for Verifying String Hypothesis in Continuous Speech Recognition *II* Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, 1995, pp. 281-284
63. Sanderson C, Bengio S., Boulard H., Mariethoz J., Collobert R., BenZeghiba M. F., Cardinaux F., Marcel S. Speech and Face Based Biometric Authentification at IDIAP *II* Proceedings of the International Conference on Miltimedia and Expo , 2003, pp. 1-4
64. San-Segundo R., Pellom B., Hacioglu K.,Ward W. Confidence Measures for Spoken Dialogue Systems *II* Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, 2001, pp. 393-396
65. Schaaf T., Kemp T. Confidence Measures for Spontaneous Speech Recognition *II* Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, 1997, pp. 875-878

66. Sigurdsson S., Peterson K. B., Lehn-Schioler T. Mel Frequency Cepstral Coefficients: An Evaluation of Robustness of MP3 Encoded Music *II* Proceedings of the International Conference on Music Information Retrieval, 2006, pp. 286-289
67. Siu M. H., Mark B., Au W. H. Minimization of Utterance Verification Error Rate as a Constrained Optimization Problem *II* IEEE Signal Processing Letters, 2006, v. 13, pp. 760-763
68. Siu M., Gish H. Evaluation of Word Confidence for Speech Recognition Systems *II* Computer Speech And Language, 1999, pp. 299-319
69. Soong F. K., Lo W. K. Generalized Posterior Probability for Minimum Error Verification of Recognized Sentences *II* Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, 2005, pp. 85-88
70. Stephenson T. A., Boulard H., Bengio S., Morris A. C. Automatic Speech Recognition Using Dynamic Bayesian Networks with Both Acoustic and Articulatory Variables *II* Proceedings of the International Conference on Spoken Language Processing, 2000, pp. 951-954
71. Sukkar R. A. Rejection for Connected Digit Recognition Based on GPD Segmental Discrimination *II* Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, 1994, pp. 393-396
72. Sukkar R. A., Lee C. H. Vocabulary Independent Discriminative Utterance Verification for Nonkeyword Rejection in Subword Based Speech Recognition *II* IEEE Transactions on Speech and Audio Process, 1996, V. 4, pp. 420-429
73. Uhrik C, Ward W. Confidence Metrics Based on N-gram Language Model Back-off Behaviors *II* Proceedings of the European Conference on Speech Communication and Technology, 1997, pp. 2772-277'4

74. Ullman J. D., Hopcroft J. E. Introduction to Automata Theory, Language and Computation *II* Addison Wesley, 1979

75. Weintraiib M., Beaufays F., Rivlin Z., Konig Y., Stolcke A. Neural-network Based Measures of Confidence for Word Recognition *II* Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, 1997, pp. 887-890

16. Weitraub M. LVCSR Log-likelihood Ratio Scoring for Keyword Spotting *II* Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, 1995, pp. 297-300

77. Wessel F. Word Posterior Probabilities for Large Vocabulary Speech Recognition *II* Ph.D. Thesis,RWTH Aachen University, German, 2002

78. Wessel F., Macherey K., Ney H. A Comparison of Word Graph and N-Best List Based Confidence Measures *II* Proceedings of the European Conference on Speech Communication and Technology, 1999, pp. 315-318

79. Wessel F., Macherey K., Schluter R. Using Word Probabilities as Confidence Measures *II* Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, 1998, pp. 225-228

80. Wessel F., Schluter R., Macherey K., Ney H. Confidence Measures for Large Vocabulary Continuous Speech Recognition *II* IEEE Transactions on Speech and Audio Process, 2001, pp. 288-298

81. Wessel F., Schluter R., Ney H. Using Posterior Word Probabilities for Improved Speech Recognition *II* Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, 2000, pp. 1587-1590

82. Young S. J. A Review of Large-Vocabulary Continuous Speech Recognition *II* IEEE Signal Processing Magazine, 1996, pp. 45-57

83. Young S., Evermann G., Hain T. Kershaw D., Moore G., Odell J., Ollason D., Povey D., Valtchev V., Woodland P. The HTK Book *II* Cambridge University Engineering Department, 2002

84. Zhang R., Rudnicky A. I. Word Level Confidence Annotation Using Combinations of Features *II* Proceedings of the European Conference on Speech Communication and Technology, 2001, pp. 2105-2108
85. Zweig G. Speech Recognition with Dynamic Bayesian Networks *II* Ph.D. Thesis, University of California, Berkeley 1998