УДК 004.934.2

О ПРОБЛЕМЕ ПАРАМЕТРИЗАЦИИ РЕЧЕВОГО СИГНАЛА В СОВРЕМЕННЫХ СИСТЕМАХ РАСПОЗНАВАНИЯ РЕЧИ

Т.В.Шарий

Введение.

На протяжении многих лет развиваются системы машинного синтеза и распознавания речи. Совместное использование таких систем является фундаментом полноценного голосового интерфейса, спектр применения которого на практике чрезвычайно широк. Исследованиями в области речевого интерфейса занимаются многие ученые, а разработки ведут крупнейшие компьютерные организации, в том числе Intel и IBM.

Конечной целью создания автоматических систем распознавания речи (АСРР) является способность машины распознавать слова в акустическом сигнале с эффективностью, не меньшей по сравнению с аналогичной способностью человека. В ходе истории разработок АСРР наблюдался значительный прогресс. Размер словаря вырос до нескольких миллионов слов, а сами системы эволюционировали от дикторозависимых к дикторонезависимым. Тем не менее, главные проблемы на сегодняшний день не решены. Эти проблемы связаны с вариабельностью речи и вызваны искажением речи фоновым шумом, явлением коартикуляции, а также зависимостью речевых характеристик от голоса и интонации [1]. В таблице 1 приведены данные сравнительного анализа эффективности распознавания отдельных слов человеком и компьютером в условиях отсутствия шума [2].

Таблица 1. Сравнение показателей WER (ошибки распознавания слов) человека и машины для разных видов речи

Вид представления речи	**Человек (%)**	**Машина (%)**
Связанные разряды (Connected digits)	0.009	0.72
Транзактная модель (Transactional RM)	0.1	3.6
Диктовка текста (Wall Street Journal)	0.9	7.2
Изолированные буквы	1.6	5.0
Произвольная телефонная речь	4.0	43.0

Целью данной работы является обзор и сравнительный анализ наиболее популярных подходов к эффективной параметризации речевых сигналов в современных АСРР.

Архитектура современных систем распознавания речи. Архитектура современных АСРР (рис.1) включает два основных модуля – модуль предобработки сигнала (front-end) и модуль постобработки сигнала (back-end) [1].

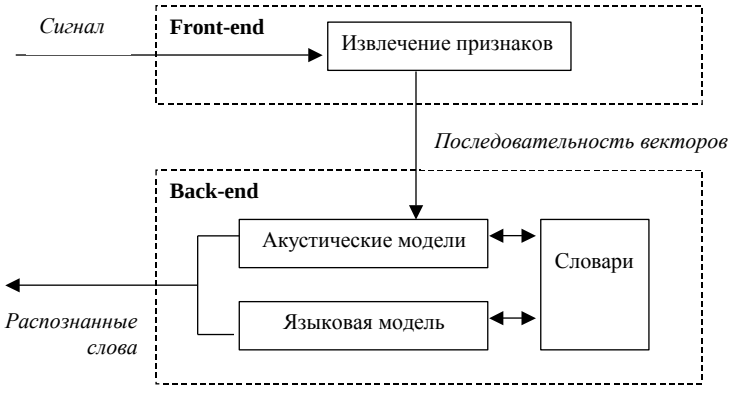

Рис. 1. Архитектура типовой АСРР.

Модуль предобработки выполняет захват речевого сигнала и его цифровую обработку. На выходе получается последовательность векторов признаков речевого сигнала. Вектор признаков представляет собой некоторое компактное описание сигнала, позволяющее при этом максимизировать показатели эффективности распознавания. Процесс получения векторов признаков называется извлечением признаков. Модуль постобработки сигнала выполняет непосредственно распознавание последовательности слов в речевом сигнале на основе закона условных вероятностей Байеса [3]. В этом модуле реализованы

акустическая и *языковая* модели речи для конкретной области применения системы. Языковая модель позволяет получить значение вероятности появления какой-либо последовательности слов независимо от наблюдаемой последовательности. Акустическая модель рассчитывает вероятность, с которой наблюдаемая последовательность слов является конкретной заданной последовательностью.

Современные АСРР основываются на статистическом моделировании. Эти системы сначала обучаются на многочасовых коллекциях речевых данных (процесс обучения заключается в настройке параметров статистических моделей). Затем, на этапе распознавания, системы производят сопоставление входных образов с ранее вводившимися по обученным моделям. Такой подход обладает существенным недостатком – он эффективен, только если речевые характеристики обучающих образов и тестовых близки, чего на практике добиться сложно. Тем не менее, на сегодняшний день статистический метод Скрытых Марковских моделей (HMM) является стандартом постобработки речевых сигналов [1,3].

Задача параметризации речевого сигнала стоит наиболее остро и до сих пор не решена в полной мере, а именно: не найдено адекватной модели представления речевых структурных единиц, позволившей бы с высокой точностью распознавать речь. Эффективность модуля постобработки сигнала напрямую зависит от качества извлечения признаков. К основным методам параметризации можно отнести: линейное предсказание, кепстральный анализ, вейвлет-преобразование, анализ спектра модуляции [4].

Методы MFCC, LPCC, PLP. Большинство современных АСРР сосредотачивают усилия на извлечении частотной характеристики речевого тракта человека, отбрасывая при этом характеристики сигнала возбуждения. Это объяснено тем, что коэффициенты первой модели обеспечивают лучшую разделимость звуков.

Для отделения сигнала возбуждения от сигнала речевого тракта прибегают к кепстральному анализу. Схематически этот метод представлен на рис.2:

Рис. 2. Общая схема кепстрального анализа сигнала

где FFT – блок быстрого преобразования Фурье сигнала (БПФ), LOG – блок логарифмирования спектра, IFFT – блок обратного быстрого преобразования Фурье (ОБПФ).

К самым мощным методам, основанным на кепстральном анализе сигнала, относятся: метод кепстральных коэффициентов линейного предсказания (LPCC) [1], метод коэффициентов перцептивного линейного предсказания (PLP) и робастный PLP (PLP-RASTA) [5], метод кепстральных коэффициентов на шкале мел (MFCC) [6].

Первые два этапа цифровой обработки сигнала одинаковы для перечисленных методов. Это предварительное усиление (pre-emphasis) и сегментация на фреймы. На первом этапе к сигналу применяется БИХ-фильтр вида:

$$H_{pre}(z) = 1 + a_{pre} z^{-1} \tag{1}$$

Данный фильтр позволяет «усилить» высокочастотную область спектра сигнала. Во-первых, это нужно для выравнивания спектра, т.к. вокализованные участки речи характеризуются резко спадающим спектром. Во-вторых, человеком лучше воспринимаются частоты выше 1кГц. Значение коэффициента a_{pre} обычно выбирается из промежутка [-1.0, -0.4].

На втором этапе речевой сигнал разбивается во времени на перекрывающиеся короткие промежутки (фреймы), в которых проводится «мгновенный» кепстральный анализ. Обычно продолжительность фрейма составляет от 20мс до 40мс. Полагается, что на этих участках речевой сигнал можно считать квазистационарным. К фрейму применяется оконная функция Хемминга:

$$w(n) = 0.54 - 0.46 \cos\left(\frac{2\pi n}{N}\right) \tag{2}$$

Ниже приводится краткое описание методов параметризации сигнала, основанных на кепстральном анализе.

1. Алгоритм *LPCC* [1] начинается с вычисления p коэффициентов $\{a_k\}_{k=1}^{p}$ авторегрессионной модели для каждого фрейма на основе модели $\hat{S}$:

$$\hat{S}(z) = \frac{A}{1 - \sum_{k=1}^{p} a_k z^{-k}} \tag{3}$$

После того, как все параметры модели найдены, вычисляются кепстральные LPCC-коэффициенты по рекурсивной функции:

$$c(n) = \begin{cases} 0 & n < 0 \\ \log_{\varepsilon}(A) & n = 0 \\ a_n + \sum_{k=1}^{n-1} \left(\frac{k}{n}\right) c(k) a_{n-k} & 0 < n < p \\ \sum_{k=n-p}^{n-1} \left(\frac{k}{n}\right) c(k) a_{n-k} & n > p \end{cases} \tag{4}$$

На основе конечного числа коэффициентов линейного предсказания может быть получено бесконечное число LPCC-коэффициентов. Экспериментально установлено, что 12-20 коэффициентов достаточно для формирования оптимального для данного метода вектора признаков.

2. Алгоритм *PLP* [5] отличается от предыдущего тем, что учитывает особенности восприятия различных частот человеком – перед вычислением параметров авторегрессионной модели сигнал проходит определенную предобработку. Алгоритм схематически представлен на рис.3.

В блоке 1 вычисляется мгновенный спектр Фурье в текущем фрейме. В блоке 2 спектр Фурье преобразуется в спектр на шкале барков, после чего выполняется операция свертки маскирующих кривых критических полос с полученным спектром для получения эффекта маскировки частоты. В блоке 3 к данным применяется функция кривой одинаковой громкости для аппроксимации уровня чувствительности человека к слышимому звуку 40дБ. В блоке 4, исходя из закона восприятия громкости звука человеком, из спектральных коэффициентов извлекается кубический корень. Блоки 5 и 6 вычисляют значение выражений (3) и (4) соответственно.

Рис. 3. Схематическое описание алгоритма PLP

Преимуществом метода PLP по сравнению с LPCC является то, что он позволяет подавить информацию, связанную с индивидуальными характеристиками диктора, путем выбора подходящего порядка модели. Тем не менее, данный метод более чувствителен к частоте основного тона.

3. Алгоритм *MFCC* [6] не уступает в эффективности алгоритму PLP, при этом является гораздо более простым в реализации. Алгоритм схематически представлен на рис.4.

Рис. 4. Схематическое описание алгоритма MFCC

В блоке 1 вычисляются коэффициенты спектра Фурье $\{X_k\}_{k=0}^{N-1}$.

В блоке 2 на вычисленный спектр накладывается набор из M фильтров шкалы мел (обычно M=20 или M=24):

$$x_i = \sum_{k=0}^{N-1} |X_k| \cdot H_i(f_k), \quad i = 1..M \tag{5}$$

Фильтр шкалы мел *H* имеет треугольный вид:

$$H_i(f_k) = \begin{cases} 0 & f_k < f_{c_{i-1}} \\ \dfrac{f_k - f_{c_{i-1}}}{f_{c_i} - f_{c_{i-1}}} & f_{c_{i-1}} < f_k < f_{c_i} \\ \dfrac{f_{c_{i+1}} - f_k}{f_{c_{i+1}} - f_{c_i}} & f_{c_i} < f_k < f_{c_{i+1}} \\ 0 & f_k > f_{c_{i+1}} \end{cases} \tag{6}$$

В формуле (6) значения f_{c_i} рассчитываются исходя из центральных мел-частот:

$$f_{c_i} = 700 \cdot (e^{\hat{f}_{c_i}/1127} - 1) \tag{7}$$

В блоке 3 выполняется логарифмирование измененного спектра:

$$x_i = \log(x_i), \quad i = 1..M \tag{8}$$

Благодаря логарифмированию достигается эффективное сжатие пространства признаков и преимущества гомоморфной обработки. Однако логарифм малых чисел стремится к минусу бесконечности. Чтобы обойти этот эффект, можно применить метод маскировки (log(x+c)) либо заменить логарифм кубическим корнем (и то, и другое приводит к снижению качества распознавания).

В блоке 4 производится дискретное косинусное преобразование:

$$c_j = \sum_{i=1}^{M} x_i \cos(j \cdot (i - 0.5) \cdot \frac{\pi}{M}), \quad j = 1..J \tag{9}$$

Обычно число MFCC-коэффициентов J для формирования вектора признаков выбирают равным 12. Наиболее релевантная информация содержится в первых 6 коэффициентах. Важность включения остальных коэффициентов определяется конкретным случаем и диктором.

Сравнительный анализ показателей эффективности описанных методов параметризации на речевой базе Aurora 2.0 приведен в таблице 2 [7].

Таблица 2. Сравнение показателей WAR (процент правильно распознанных слов) наиболее эффективных методовпараметризации

	Обучение без шума			**Обучение с шумом**		
	0dB	**1-20dB**	**-5dB**	**0dB**	**1-20dB**	**-5dB**
LPCC	99.13	87.21	23.92	99.41	56.50	3.61
MFCC	99.21	88.25	22.27	99.65	54.28	7.23
PLP	99.40	89.81	28.53	99.65	62.05	6.98
MSG	97.42	86.37	18.05	98.90	56.06	2.19

Видно, что качественно методы, основанные на кепстральной обработке сигнала, практически не отличаются (небольшие расхождения в показателях вызваны, скорее, спецификой речевой базы). Поэтому выбор метода остается за предпочтениями исследователя. В таблице 2 приведен также эффективный метод, не основанный на кепстральном анализе, – анализ спектра модуляции (MSG). Видно, что данный метод уступает любому из «кепстральных» методов.

Часто для отражения временных изменений в вектор признаков помимо самих коэффициентов добавляют их первые и вторые производные (дельта-характеристики). Первая производная вычисляется в соответствии с выражением:

$$\partial c_t(n) = \frac{\sum_{k=-K}^{K} k c_{t+k}(n)}{\sum_{k=-K}^{K} k^2}, \quad K > 0 \tag{10}$$

Включение производных в вектор признаков позволяет также снизить влияние сверточных искажений сигнала, в силу того, что эти искажения обычно медленно изменяются во времени и аддитивны в кепстральной области.

Робастное извлечение признаков. Для снижения влияния произвольности речи и фонового шума в АСРР исследователями были предложены методы робастного извлечения признаков. К наиболее эффективным относятся следующие методы [8]: нормализация кепстрального среднего, маскировка шума, RASTA-обработка.

Суть метода *нормализации кепстрального среднего* (CMN), известного также как метод вычитания кепстрального среднего (CMS), заключается в попытке удалить «сверточные» искажения в сигнале путем вычитания из кепстральных характеристик их среднего значения (математического ожидания):

$$\bar{c}_t = c_t - \frac{1}{T}\sum_{i=0}^{T-1} c_i \tag{11}$$

Любое стационарное спектральное искажение, например, частотная характеристика микрофона, таким образом, удаляется в связи с тем, что эффекты в области свертки аддитивны в кепстральной области. Кроме того, экспериментально показано, что метод CMN также уменьшает чувствительность к голосу. Теоретически это можно объяснить тем, что информация о дикторе может быть представлена не только в постоянной компоненте спектра, а и распределена на низких частотах спектра модуляции.

Метод *маскировки шума* заключается в добавлении некоторой константы C к спектральным коэффициентам при вычислении кепстра:

$$c = DCT(\log(C + x^{e^{(j\omega)}})) \tag{12}$$

Таким образом, метод не учитывает спектр шума, а целью его является достижение устойчивости кепстра к изменениям шума. Недостатком данного метода является то, что с ростом уровня маскировки эффективность падает.

Суть метода *RASTA-обработки* признаков заключается в удалении незначащих компонент лингвистического сообщения путем применения полосовых фильтров ко временны*м* траекториям векторов признаков между двумя нелинейными операциями (например логарифмированием). С помощью такой фильтрации RASTA не только удаляет постоянную компоненту спектра, как это делает метод CMN, а еще и изменяет сам спектр сигнала.

Обычно при этом подходе применяется БИХ-фильтр типа:

$$H(z) = 0.1z^4\left(\frac{2 + z^{-1} - z^{-3} - 2z^{-4}}{1 - 0.98z^{-1}}\right) \tag{13}$$

Этот фильтр позволяет лучше разграничить соседние звуки речи.

Интересным представляется *метод MVA*, предложенный в [7]. Данный метод включает в общем случае три этапа:

1) вычитание математического ожидания (среднего) вектора признаков. Эта операция выполняется аналогично (11);
2) нормализация дисперсии:

$$\hat{c}_t[j] = (\sigma^2)^{1/2} \cdot \bar{c}_t[j] \tag{14}$$

$$\sigma^2[j] = \frac{1}{T}\sum_{t=1}^{T}(c_t[j] - \frac{1}{T}\sum_{t=1}^{T} c_t[j])^2 \tag{15}$$

3) ARMA-фильтрация (авторегрессия и скользящее среднее):

$$\breve{c}_t = \frac{\breve{c}_{t-m} + \ldots + \breve{c}_{t-1} + \hat{c}_t + \ldots + \hat{c}_{t+m}}{2m+1} \tag{16}$$

В методе MVA последовательно обрабатываются фреймы – текущий вектор признаков модифицируется в соответствии с предыдущими согласно (14)-(16), а индекс t обозначает номер фрейма

в последовательности фреймов. Метод имеет малую вычислительную сложность и при этом является более эффективным по сравнению с методами, описанными выше.

Кодирование последовательности векторов признаков. Временн*ая* информация о сигнале может быть закодирована не только с помощью дельта-характеристик. Для компактного описания отдельных речевых фрагментов (фонем, дифтонгов, слогов) могут применяться методы обработки последовательности векторов признаков, такие как: метод кепстрально-временных матриц (CTM) [9], дискретное косинусное преобразование (DCT), дискретное преобразование Лежандра (DLT), преобразование Карунена-Лоэва (KLT) [8]. Показано, что из этих методов KLT обеспечивает лучшие результаты распознавания. Однако на практике эти методы используются реже, чем дельта-характеристики, т.к. для них критичен вопрос автоматической сегментации сигнала.

Заключение. Проблема параметризации речевого сигнала в контексте создания АСРР актуальна и нуждается в решении. Как видно из анализа, приведенного в статье, при использовании современных методов параметризации речи, процент верно распознаваемых слов, колеблется в широком диапазоне от 20% до 99%. Этого явно недостаточно для создания эффективных АСРР, в которых максимально допустимая ошибка распознавания не должна превышать 2%. Применение робастных методов извлечения признаков, а также учет модели речевого тракта человека при первичной обработке сигнала, позволяет заметно повысить качество распознавания. Отмечено, что методы, основанные на кепстральном анализе сигнала, наиболее эффективны.

Одним из направлений усовершенствования существующих методов параметризации речевых сигналов является динамическая настройка размерности вектора признаков. При современных подходах размерность пространства признаков устанавливается заранее, на основе экспериментальных данных. В то же время имеет смысл адаптивно настраивать размерность вектора признаков на основе анализа главных компонент (KLT), т.к. в некоторых случаях старшие коэффициенты могут «зашумлять» пространство признаков. При этом дельта-характеристики также должны быть сохранены.

РЕЗЮМЕ

У статті розглядаються принципи побудови сучасних автоматичних систем розпізнавання мови. Проводиться порівняльний аналіз найбільш популярних методів параметризації мовних сигналів. Описані методи робастного отримання ознак в умовах шуму.

SUMMARY

The article considers the principles of the state-of-the-art automatic speech recognition systems architecture. The comparative analysis of the most popular speech signal parameterization techniques is given. The robust feature extraction techniques in the noise conditions are reviewed.

СПИСОК ЛИТЕРАТУРЫ

1. X.Huang, A.Acero, H.Hon. Spoken Language Processing: A guide to theory, algorithm, and system development. Prentice Hall, 2001.
2. R.P.Lippmann. Speech recognition by machines and humans. Speech communications, 22:1–16, 1997.
3. C.Lee. On Automatic Speech Recognition at the Dawn of the 21st Century. IEICE TRANS.Fundamentals, p.1-20, 2001.
4. S.Sukittanon, L.E.Atlas, J.W.Pitton, “Modulation-Scale Analysis For Content Identification”, IEEE Transactions On Signal Processing, Vol.52, No.10, Oct. 2004: pp.3023-3035.
5. H.Hermansky. Perceptual Linear Predictive (PLP) Analysis of Speech. The Journal of the Acoustical Society of America, 87(4): 1738-1752, 1990.
6. F.Zheng, G.Zhang, Z.Song. Comparison Of Different Implementations Of MFCC. J. Computer Science & Technology, 16(6): 582-589, 2001.
7. C.-P.Chen, J.Bilmes. MVA Processing of Speech Features. Technical Report UWEETR-2003-0024, University of Washington, Dept. Of EE, 2003.
8. B.Milner. A Comparison of Front-End Configurations for Robust Speech Recognition. In International Conference on Acoustics, Speech and Signal Processing, volume 1, pp.797-800, 2002.
9. B.Milner. Inclusion of Temporal Information into Features for Speech Recognition. In International Conference on Speech and Signal Processing, pp.256-259, 1996.

Надійшла до редакції 28.11.2008 р.