УДК 004.934.1

А. В. ТКАЧЕНЯ

К ВОПРОСУ ОБ УСТОЙЧИВЫХ К ЭМОЦИЯМ ИНФОРМАТИВНЫХ ПРИЗНАКАХ ДЛЯ ЗАДАЧИ РАСПОЗНАВАНИЯ РЕЧИ

Описан метод параметризации речевого сигнала, который дает устойчивый к эмоциям и инвариантный к диктору информативный признак (ИП) на основе кепстральных коэффициентов, определенных на экспоненциально-логарифмической шкале частот, для спектра, рассчитанного по параметрам линейного предсказания (ЛПСКК). При помощи полученного информативного признака решается задача распознавания эмоциональной речи на основе скрытых марковских моделей. В экспериментальной части показано, что использование линейного предсказания (ЛП) для вычисления спектра речевого сигнала более предпочтительно по сравнению с быстрым преобразованием Фурье, так как влияние изменения частоты основного тона на качество анализа линейного предсказания практически отсутствует, что обусловливает высокую эффективность распознавания гласных звуков с различной эмоциональной окраской. Также приводятся доводы в пользу применения экспоненциально-логарифмической шкалы частот при расчете кепстральных коэффициентов, так как ее использование позволяет снизить изменчивость пространства признаков. Полученные в ходе эксперимента результаты свидетельствуют о том, что рассматриваемый в статье информативный признак позволяет повысить эффективность распознавания эмоциональной речи на 4 %.

Ключевые слова: распознавание эмоциональной речи; информативный признак; коэффициенты линейного предсказания; кепстральные коэффициенты; скрытые марковские модели.

This paper presents a method of the speech signal parameterization that gives a robust to emotions and invariant to a speaker feature vector. For this purpose the cepstral coefficients based on a linear prediction power spectrum defined on the ExpoLog frequency scale is used. The described feature vector is applied for emotional speech recognition based on hidden Markov models. It is shown that usage of linear prediction to calculate the spectrum of a speech signal is more preferably compared with fast Fourier transform as the effect of a change in the pitch frequency on the quality of a linear predictive analysis is very insignificant, and a high recognition efficiency of vowel sounds with different emotional coloring is offered. Also, it is proposed to use the ExpoLog frequency scale in calculations of cepstral coefficients owing to the reduced variability of the feature vector space. The experimental results obtained show that usage of the described feature vector contributes to the improved (by 4 %) efficiency of emotional speech recognition.

Key words: emotional speech recognition; feature vector; linear prediction coefficients; cepstral coefficients; hidden Markov models.

Известно, что снижение эффективности распознавания речи связано с несоответствием акустических характеристик обучающих и тестируемых данных. Как показывает исследование [1], эффективность распознавания эмоциональной речи, по сравнению с нейтральной, снижается от 20 до 60 %.

В работе [2] было показано, что в зависимости от типа эмоции частота основного тона, длительность, интенсивность и речевой тракт меняются различным образом. Это свидетельствует о том, что спектральная структура речи изменяется при различных эмоциях, что приводит к изменению пространства признаков.

Для улучшения эффективности распознавания эмоциональной речи в литературных источниках было предложено несколько методов. Все их можно разделить на 3 категории: использование устойчивых информативных признаков (ИП); методы компенсации эмоций в ИП; методы адаптации моделей. Два последних метода предполагают наличие дополнительного этапа анализа базы эмоциональной речи, который необходим для моделирования статистических данных о каждой из эмоций с последующим их включением в систему распознавания речи. Главный недостаток этих двух методов заключается в необходимости определения эмоций в тестовой выборке для применения соответствующей компенсации эмоции или адаптированной модели.

В связи с этим в настоящей статье рассмотрен подход, основанный на использовании устойчивого ИП. В работе [3] был предложен ИП, который показал хорошие результаты при распознавании речи в состоянии стресса: «крик» (LOUD) и эмоциональное состояние – «гнев» (ANGRY).

В предлагаемом исследовании автором предпринята попытка расширить область применения ИП [3] на распознавание эмоциональной речи. Необходимые для этого предпосылки, уточнения и дополнения приводятся далее по тексту. Блок-схема формирования этого ИП, назовем его ЛПСКК, приведена на рис. 1.

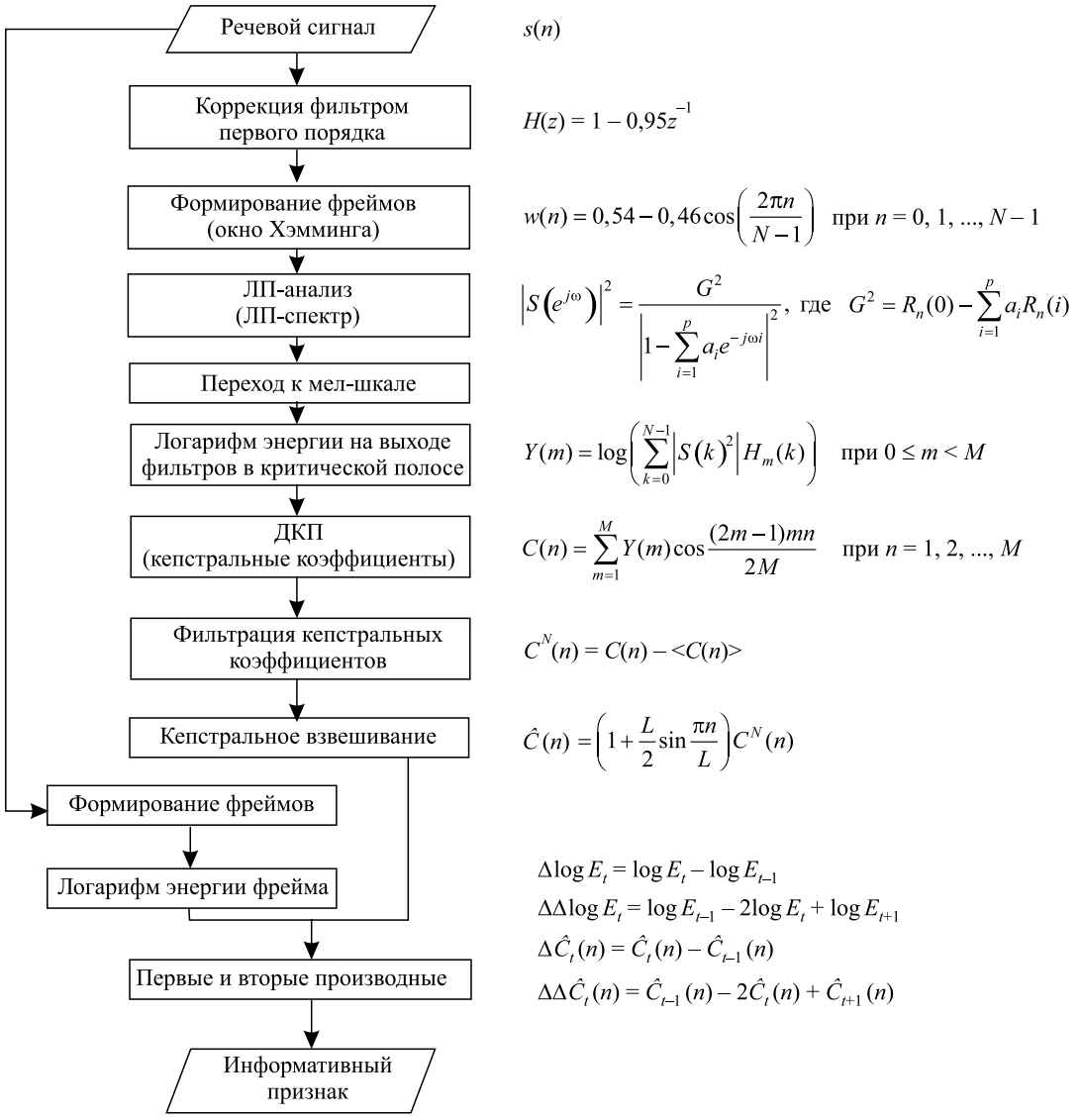

Рис. 1. Блок-схема формирования информативного признака ЛПСКК

Устойчивый к эмоциям информативный признак

ЛП-анализ (ЛП-спектр). Выбор линейного предсказания для расчета спектра мощности сигнала объясняется отсутствием влияния изменения частоты основного тона на качество ЛП-анализа, что обусловливает высокую эффективность распознавания гласных звуков, а минус заключается в отсутствии нулей в полученном спектре, что приводит к спутыванию схожих согласных [4]. Зная, что частота основного тона меняется в зависимости от типа эмоции [2], можно предположить, что эффективность распознавания эмоциональной речи при использовании ЛП-анализа для расчета спектра, по сравнению с быстрым преобразованием Фурье, будет выше.

Переход к мел-шкале. Как было показано в [3], для эмоционального состояния «гнев» 2-я форманта (диапазон частот от 1250 до 1750 Гц) меньше всего изменяется по сравнению с нейтральным эмоциональным состоянием. В связи с этим авторами статьи [3] было предложено использовать экспоненциально-логарифмическую шкалу частот (1), приведенную на рис. 2. Ее использование позволяет снизить изменчивость пространства информативных признаков, а это приводит к увеличению эффективности распознавания речи, что было подтверждено в статье [3].

$$\Theta_{\text{ExpoLog}}(f)=\begin{cases}700\left(10^{f/3988}-1\right) & \text{при } 0\le f\le 2000 \text{ Гц},\\ 2595\log_{10}\left(1+\dfrac{f}{700}\right) & \text{при } f>2000 \text{ Гц}.\end{cases} \tag{1}$$

Автором настоящей статьи было сделано предположение о возможности использования экспоненциально-логарифмической шкалы частот для повышения эффективности распознавания эмоциональной речи, обоснованность которого будет проверена экспериментально.

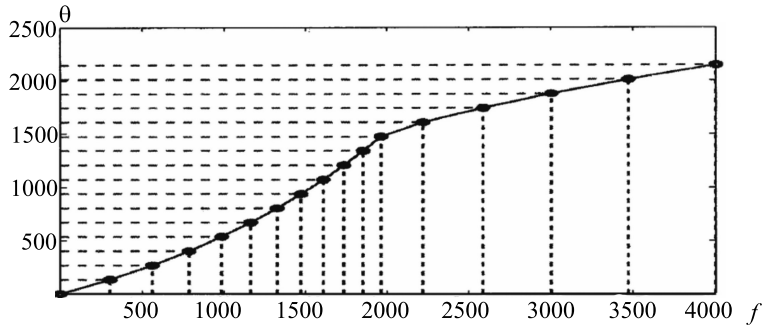

Рис. 2. Экспоненциально-логарифмическая шкала частот [3, fig. 7 (c)]

Логарифм энергии на выходе фильтров в критической полосе. Чтобы получить значения логарифма энергии на основе ЛП, необходимо найти логарифм энергии сигнала на выходе каждого из треугольных фильтров (2) [5], представленных на рис. 3.

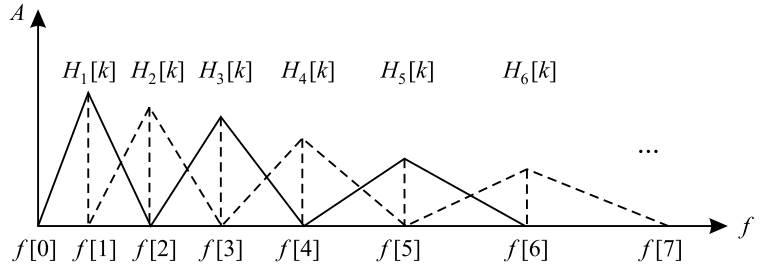

Рис. 3. Набор треугольных фильтров на шкале частот [5, fig. 6.28]

$$H_m[k]=\begin{cases}0, & f_k<f[m-1],\\ \dfrac{2\left(f_k-f[m-1]\right)}{\left(f[m+1]-f[m-1]\right)\left(f[m]-f[m-1]\right)}, & f[m-1]\le f_k\le f[m],\\ \dfrac{2\left(f[m+1]-f_k\right)}{\left(f[m+1]-f[m-1]\right)\left(f[m+1]-f[m]\right)}, & f[m]\le f_k\le f[m+1],\\ 0, & f_k>f[m+1].\end{cases} \tag{2}$$

Полоса пропускания всех фильтров (2) в экспоненциально-логарифмической шкале частот должна быть одинаковой и равняться 108 мелам, что соответствует одной критической полосе для слуха человека [4]. Тогда, зная частоту среза сигнала $f_{ср}$, можно найти значения частот $f[m]$, воспользовавшись следующей формулой:

$$f[m](\Theta) = \begin{cases} 3988\log_{10}\left(1+\dfrac{\Theta_m}{700}\right) & \text{при } 0 \le f \le 2000 \text{ Гц}, \\ 700\left(10^{\Theta_m/2595}-1\right) & \text{при } f > 2000 \text{ Гц}, \end{cases}$$

где $\Theta_m = (\Theta_{max}/M)i$ при $i = 1, 2, ..., M$; $\Theta_{max} = \Theta(f_{ср})$, которую можно рассчитать по формуле (1); $M = \Theta_{max}/108$ и $f[0] = 0$.

Дискретно-косинусное преобразование (ДКП) (кепстральные коэффициенты). В результате применения ДКП для каждого фрейма будет получено M кепстральных коэффициентов. Известно, что для кепстральных коэффициентов с высоким индексом характерно скачкообразное изменение их величины, что затрудняет создание устойчивых моделей и, как следствие, снижает эффективность распознавания речи [4]. Поэтому в формировании ИП было предложено использовать только первые 12 кепстральных коэффициентов $(c_1, c_2, ..., c_{12})$, а вместо 0-го кепстрального коэффициента взять значение логарифма энергии сигнала, полученного на соответствующем фрейме без применения коррекции сигнала (см. левую часть на рис. 1). Таким образом, будет сформировано пространство признаков с размерностью, равной 13.

Фильтрация кепстральных коэффициентов. В работе [4] указано, что оценка среднего значения кепстральных коэффициентов по всему высказыванию позволяет в значительной степени снизить влияние индивидуальных особенностей голоса диктора (т. е. параметров голосового тракта). Проведенные в [6, 7] исследования, направленные на нормализацию кепстральных коэффициентов для речи с различными стилями произношения, также показали хорошие результаты, что свидетельствует о возможности повысить эффективность распознавания эмоциональной речи при использовании нормализации ИП, так как в эмоциональной речи могут сильно проявляться индивидуальные особенности голоса диктора.

Чтобы осуществить нормализацию кепстральных коэффициентов, необходимо определить среднее значение величины кепстральных коэффициентов для всех фреймов, входящих в выбранный фрагмент речевых данных ($<c_1>$, $<c_2>$, ..., $<c_{12}>$), а затем вычесть их из соответствующих значений величин кепстральных коэффициентов выбранного кепстра, для которого осуществляется нормализация.

Однако вычисление среднего на всем высказывании приводит к большой задержке распознавания речи. Для того чтобы этого избежать, применяется фильтрация ИП, которая отличается от нормализации тем, что среднее значение величины кепстральных коэффициентов определяется не на всей длине речевого сообщения, а на фрагментах постоянной длительности T:

$$-\frac{T-1}{2} \le n \le \frac{T-1}{2}. \qquad (3)$$

Как видно из (3), усреднение должно выполняться на половине предшествующих и половине будущих фреймов. В ходе экспериментов была определена оптимальная длительность фрагмента усреднения, равная 5 с. Она обеспечивает наилучшее отношение увеличения эффективности распознавания речи к минимальной задержке, которая для случая фрагмента усреднения в 5 с составляет не менее 2,5 с.

Первые и вторые производные. В начале 1980-х гг. Фуруи [8], желая использовать информацию о динамике речи для верификации распознанных результатов, вместе с исходными кепстральными коэффициентами вводит в информативный признак параметры, характеризующие спектральные переходы. В качестве динамических параметров речи хорошо себя зарекомендовали первые и вторые производные кепстральных коэффициентов и логарифма энергии во фрейме [9].

В результате сформирован ИП, который можно считать устойчивым к эмоциям и инвариантным к диктору, а его размерность пространства признаков будет равна 39 (логарифм энергия во фрейме + 12 кепстральных коэффициентов + 13 первых и 13 вторых производных).

База эмоциональной речи и система распознавания

Для сравнительного анализа эффективности распознавания эмоциональной речи будет использована база данных японской эмоциональной речи университета Кейо (Keio-ESD) [10], так как в базе представлен широкий набор различных эмоциональных состояний. База состоит из 20 слов, каждое из которых произносится один раз с 47 различными эмоциями, исключение составляет только нейтральное эмоциональное состояние, для которого сохранены минимум 4 варианта произношения для каждого из слов. Запись всех слов осуществляется одним диктором – мужчиной 32 лет, частота дискретизации сигнала 16 000 Гц, разрядность квантования 16 б, формат звуковых файлов – Waveform Audio File ormat (WAV).

Обучение и тестирование проводились на основе перекрестной проверки (k-fold cross-validation [11]) с разбиением нейтральных речевых данных на 10 равных частей. При этом обучающая выборка должна состоять из нейтральных и эмоциональных речевых данных в соотношении 1 : 1. Тестовая выборка представляет собой последовательность слов из базы Keio-ESD, разделенных тишиной.

Система распознавания речи реализована на основе скрытых марковских моделей (СММ) при помощи программного пакета HTK [12]. Значения параметров СММ оцениваются на обучающей выборке с учетом ее ручной транскрипции по фонемам. Результат распознавания получается путем выбора последовательности слов с максимальной апостериорной вероятностью.

Для определения величины эффективности распознавания эмоциональной речи была применена следующая формула:

$$WAcc = \frac{N - S - D - I}{N},$$

где N – число слов в распознаваемой речи (правильная транскрипция); S – количество замененных слов в речи при распознавании; D – численность удаленных слов из речи при распознавании; I – количество слов, вставленных в речь, при распознавании.

Сравнительный анализ распознавания эмоциональной речи

Оценим эффективность описанного в статье ИП (ЛПСКК) в целом и относительного увеличения эффективности распознавания эмоциональной речи при использовании некоторых из блоков, представленных на рис. 1. Кроме того, сравним полученные результаты с таковыми для «стандартного» ИП (используемого по умолчанию в большинстве систем распознавания речи), который можно сформировать на основе описанного в статье ИП, заменив на рис. 1 вычисление ЛП-спектра и экспоненциально-логарифмическую шкалу частот на быстрое преобразование Фурье (БПФ) и мел-частотную шкалу соответственно. В литературных источниках такой информативный признак носит название мел-частотные кепстральные коэффициенты (МЧКК, Mel-Frequency Cepstrum Coefficients – MFCC [5]) (таблица).

В таблице используются следующие обозначения: коррекция фильтром первого порядка «FxdP», фильтрация кепстральных коэффициентов «CMFilt», кепстральное взвешивание «CepLift» и добавление первых и вторых производных «Δ», использование экспоненциально-логарифмической шкалы частот «ExpoLog» или мел-частотной шкалы «MelFreq». Результаты распознавания эмоциональной речи приводятся для трех случаев: эффективность распознавания речи с нейтральным эмоциональным состоянием «Нейтр.», эмоциональным состоянием «Гнев» и эффективность распознавания речи для всех эмоциональных состояний «Общ.», представленных в базе Keio-ESD.

Эффективность распознавания эмоциональной речи для ЛПСКК и МЧКК при обучении на нейтральных и эмоциональных речевых данных в соотношении 1 : 1

Информативный признак	Спектр на основе					
	ЛП-анализа			БПФ		
	Нейтр., %	Гнев, %	Общ., %	Нейтр., %	Гнев, %	Общ., %
	MelFreq					
FxdP	71	59	62	73	58	61
FxdP + CMFilt	77	68	70	78	67	69
FxdP + CMFilt + CepLift	79	72	73	79	69	71
FxdP + CMFilt + CepLift + Δ	83	78	80	84	76	78
	ExpoLog					
FxdP	68	64	66	68	51	53
FxdP + CMFilt	75	73	76	71	58	59
FxdP + CMFilt + CepLift	77	74	77	72	58	60
FxdP + CMFilt + CepLift + Δ	83	80	82	78	62	67

Из таблицы видно, что эффективность распознавания нейтральной речи для МЧКК (верхний правый квадрант) выше, чем для ЛПСКК (нижний левый квадрант), в то время как для эмоциональной речи – наоборот. Это подтверждает сделанное ранее предположение о том, что использование ЛП-анализа позволяет добиться улучшения эффективности распознавания гласных звуков с различной эмоциональной окраской. С другой стороны, как и было сказано в [4], применение ЛП-анализа приводит к спутыванию схожих согласных, что сказывается на эффективности распознавания нейтральной речи. Результаты эксперимента свидетельствуют о том, что применение коррекции фильтром первого порядка, фильтрации кепстральных коэффициентов, кепстрального взвешивания, а также добавление первых и вторых производных позволяют повысить эффективность распознавания эмоциональной речи. Сочетание ЛП-анализа с экспоненциально-логарифмической шкалой частот (ЛПСКК) приводит к максимальной эффективности распознавания.

Таким образом, информативный признак, устойчивый к эмоциям и инвариантный к диктору, сформированный на основе кепстральных коэффициентов, определенных на экспоненциально-логарифмической шкале частот для спектра, рассчитанного по параметрам линейного предсказания, позволяет повысить эффективность распознавания эмоциональной речи на 4 % по сравнению с аналогичным информативным признаком на основе мел-частотных кепстральных коэффициентов, полученных на мел-частотной шкале для спектра, рассчитанного при помощи быстрого преобразования Фурье.

БИБЛИОГРАФИЧЕСКИЙ СПИСОК

1. Vlasenko B., Prylipko D., Wendemuth A. Towards Robust Spontaneous Speech Recognition with Emotional Speech Adapted Acoustic Models : 35th German Conference on Artificial Intelligence. 2012. P. 103–107.

2. Williams C. E., Stevens K. N. Emotions and speech: Some acoustical correlates // J. Acoust. Soc. Amer. 1972. № 52. P. 1238–1250.

3. Bou-Ghazale S. E., Hansen J. H. L. A Comparative Study of Traditional and Newly Proposed Features for Recognition of Speech Under Stress // IEEE Transactions on Speech and Audio Processing. 2000. Vol. 8, № 4. P. 429–442.

4. Рылов А. С. Анализ речи в распознающих системах. Минск, 2003.

5. Xuedong Huang, Alex Acero, Hsiao-Wuen Hon. Spoken language processing. Prentice-Hall Inc.: Upper Saddle River. New Jersey, 2001.

6. Chen Y. Cepstral domain stress compensation for robust speech recognition // Proc. IEEE Inter. Conf. on Acoustics: Speech and Signal Processing. Dallas, 1987. P. 717–720.

7. Hansen J. H. L., Bria O. N. Lombard effect compensation for robust automatic speech recognition in noise : Japan Inter. Conf. Spoken Language Processing. Kobe, 1990. P. 1125–1128.

8. Furui S. Cepstral analysis technique for automatic speaker verification // IEEE Transactions on Acoustics: Speech and Signal Processing. 1981. Vol. 29, № 2. P. 254–272.

9. Beulen K., Welling L., Ney H. Experiments with linear feature extraction in speech recognition // Speech communication and technology: European conf., ESCA. Madrid, 1995. № 2. P. 1415–1418.

10. Keio-ESD. Keio University Japanese Emotional Speech Database [Electronic resource]. URL: http://research.nii.ac.jp/src/en/Keio-ESD.html (date of access: 30.09.2013).

11. Wikipedia. K-fold cross-validation [Electronic resource]. URL: http://en.wikipedia.org/wiki/Cross-validation_%28statistics%29 (date of access: 18.08.2014).

12. Young S., Evermann G., Gales M., Hain T., Kershaw D., Liu X., Moore G., Odell J., Ollason D., Valtchev V., Woodland P. The HTK Book (for HTK v. 3.4). Cambridge, 2006.

Поступила в редакцию 18.06.2014.

Андрей Владимирович Ткаченя – аспирант кафедры радиофизики и цифровых медиатехнологий. Научный руководитель – кандидат физико-математических наук, доцент кафедры радиофизики и цифровых медиатехнологий И. Э. Хейдоров.