УДК 004.934.1
DOI: 10.17586/0021-3454-2015-58-6-443-459

МЕТОДИКА ФОРМИРОВАНИЯ УСТОЙЧИВЫХ К ЭМОЦИЯМ ИНФОРМАТИВНЫХ ПРИЗНАКОВ ДЛЯ ЗАДАЧИ РАСПОЗНАВАНИЯ РЕЧИ

А. В. Ткаченя

Белорусский государственный университет, 220030, Минск, Республика Беларусь, E-mail: tkachenia@gmail.com

Описан метод параметризации речевого сигнала, который дает устойчивый к эмоциям и инвариантный к диктору информативный признак на основе кепстральных коэффициентов, определенных на экспоненциально-логарифмической шкале частот, для спектра, рассчитанного по параметрам линейного предсказания. При помощи полученного информативного признака решается задача распознавания эмоциональной речи на основе скрытых марковских моделей. Полученные в ходе эксперимента результаты свидетельствуют о том, что рассматриваемый информативный признак позволяет повысить эффективность распознавания эмоциональной речи на 5,9 %.

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

Введение. Известно, что снижение эффективности распознавания речи связано с несоответствием акустических характеристик обучающих и тестируемых данных. Согласно исследованию [1], эффективность распознавания эмоциональной речи по сравнению с нейтральной ниже на 20—60 %. В работе [2] было показано, что частота основного тона, длительность, интенсивность и речевой тракт зависят от типа эмоции. Изменение спектральной структуры речи при различных эмоциях приводит к изменению пространства признаков.

Для повышения эффективности распознавания эмоциональной речи предложено несколько подходов: использование устойчивых к эмоциям информативных признаков (ИП), применение методов компенсации эмоций в ИП и методов адаптации моделей. Два последних подхода 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 подходов заключается в необходимости определения эмоций в тестовой выборке для их компенсации или применения адаптированной модели.

В настоящей статье рассмотрен подход, основанный на использовании устойчивого ИП. В работе [3] был предложен ИП, который показал хорошие результаты при распознавании речи диктора в состоянии стресса: „крик“ (*loud*), эмоциональное состояние — „гнев“ (*angry*).

Цель настоящей статьи — расширить область применения ИП [3] на распознавание эмоциональной речи.

Формирование устойчивого к эмоциям информативного признака. Рассмотрим алгоритм формирования ИП, приведенный на рис. 1, и опишем влияние ИП на эффективность распознавания эмоциональной речи.

$s(n)$

$H(z)=1-0{,}95z^{-1}$

$W(n)=0{,}54-0{,}46\cos\left(\dfrac{2\pi n}{N-1}\right)$ при $n=0, 1, \ldots, N-1$

$\left|S(e^{j\omega})\right|^2 = \dfrac{G^2}{\left|1-\sum\limits_{1=1}^{p} a_i e^{-j\omega i}\right|^2}$, где $G^2 = R_n(0) - \sum\limits_{1=1}^{p} a_i R_n(i)$

$Y(m) = \log\left(\sum\limits_{k=0}^{N-1} |S(k)|^2 H_m(k)\right)$ при $0 \le m < M$

$C(n) = \sum\limits_{m=1}^{M} T(m)\cos\dfrac{(2m-1)mn}{2M}$ при $n = 1, 2, ..., M$

$C^N(n) = C(n) - < C(n) >$

$\hat{C}(n) = \left(1 + \dfrac{L}{2}\sin\dfrac{\pi n}{L}\right) C^N(n)$

$\Delta \log E_t = \log E_t - \log E_{t-1}$

$\Delta\Delta \log E_t = \log E_{t-1} - 2\log E_t + \log E_{t+1}$

$\Delta\hat{C}_t(n) = \hat{C}_t(n) - \hat{C}_{t-1}(n)$

$\Delta\Delta\hat{C}_t(n) = \hat{C}_{t-1}(n) - 2\hat{C}_t(n) + \hat{C}_{t+1}(n)$

Рис. 1

Коррекция фильтром первого порядка приводит к уменьшению динамического диапазона спектра речевого сигнала в результате сглаживания спектральных кривых. Это позволяет улучшить линейную структуру формант, что повышает качество анализа линейного предсказания. Затем речевой сигнал разбивается на фреймы продолжительностью 25 мс с 50 %-ным перекрытием фреймов, которые „взвешиваются“ окном Хэмминга.

Для расчета спектра мощности сигнала используется линейное предсказание (ЛП), поскольку изменение частоты основного тона на качество ЛП-анализа практически не влияет, что обусловливает высокую эффективность распознавания гласных звуков. Однако отсутствие нулей в полученном спектре приводит к ошибочному определению схожих согласных [4]. Зная, что частота основного тона меняется в зависимости от типа эмоции [2], можно предположить, что эффективность распознавания эмоциональной речи с помощью ЛП-анализа при расчете спектра, по сравнению с быстрым преобразованием Фурье, выше.

Как было показано в [3], для эмоционального состояния „гнев“ меньше всего, по сравнению с нейтральным эмоциональным состоянием, изменяется 2-я форманта (диапазон частот 1250—1750 Гц). В связи с этим авторами статьи [3] было предложено использовать экспоненциально-логарифмическую шкалу частот (рис. 2) для снижения изменчивости пространства информативных признаков.

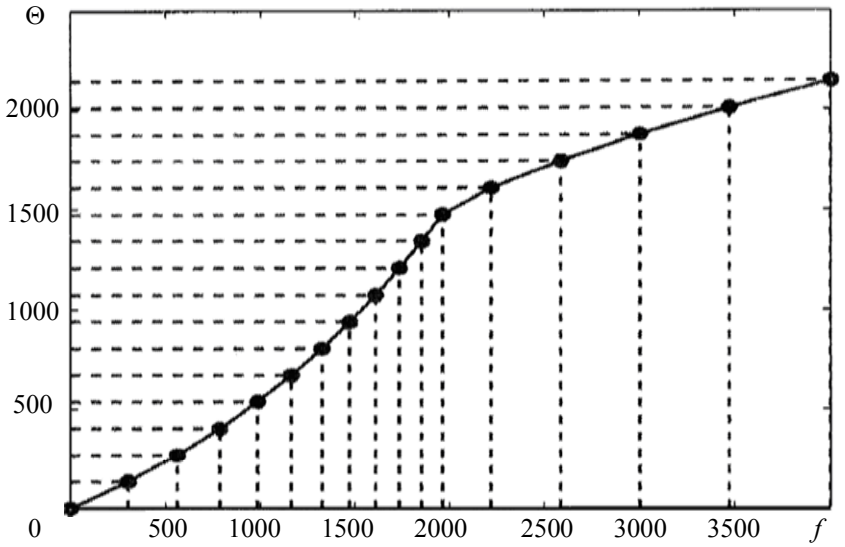

Рис. 2

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

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

Чтобы получить значения логарифма энергии на основе ЛП, необходимо найти логарифм энергии сигнала на выходе каждого из треугольных фильтров [5], представленных на рис. 3.

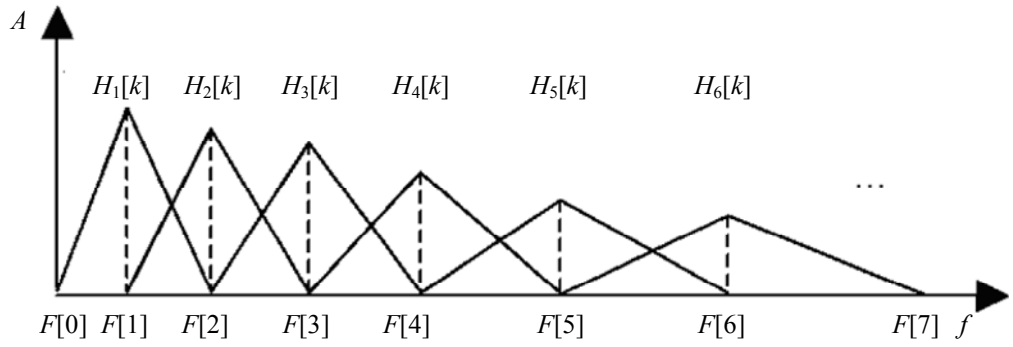

Рис. 3

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

Полоса пропускания всех фильтров (2) в экспоненциально-логарифмической шкале частот должна быть одинаковой — 108 мел, что соответствует одной критической полосе в диапазоне, воспринимаемом слухом человека [4]. Тогда, зная частоту среза сигнала $f_{\text{ср}}$, можно найти значения частот $f[m]$, воспользовавшись следующей формулой:

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

где $\Theta_m = (\Theta_{\max}/M)i$, при $i = 1,2,\ldots,M$, а $\Theta_{\max} = \Theta_{\text{ExpoLog}}(f_{\text{ср}})$, согласно формуле (1), $M = \Theta_{\max}/108$ и $f[0] = 0$.

Для вычисления кепстральных коэффициентов по логарифму энергии на выходе каждого из фильтров воспользуемся дискретно-косинусным преобразованием (ДКП). Так как фильтры (2) перекрываются на шкале частот, то значения логарифма энергии, полученные на выходе каждого из фильтров, коррелированы между собой. Согласно теории оптимального разложения функций по ортогональному базису, применение ДКП приводит к декорреляции пространства информативных признаков [6]. Это позволяет использовать диагональные ковариационные матрицы для скрытых марковских моделей (СММ), что обеспечивает снижение вычислительной сложности алгоритма распознавания эмоциональной речи.

В результате применения ДКП для каждого фрейма можно получить 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 ($c_1, c_2,\ldots, c_{12}$), а вместо нулевого взять значение логарифма энергии сигнала, полученного на соответствующем фрейме без применения коррекции сигнала (см. левую ветвь на рис. 1). Таким образом будет сформировано пространство признаков с размерностью 13.

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). Результаты проведенных в [7, 8] исследований нормализации кепстральных коэффициентов для речи с различными стилями произношения также свидетельствуют о возможности повысить эффективность распознавания эмоциональной речи за счет нормализации ИП, так как в такой речи могут проявляться индивидуальные особенности голоса диктора.

Чтобы осуществить нормализацию кепстральных коэффициентов, необходимо посчитать их средние значения для всех фреймов, входящих в выбранный фрагмент речевых данных ($<c_1>, <c_2>,\ldots, <c_{12}>$), а затем вычесть их из соответствующих значений коэффициентов выбранного кепстра.

Однако вычисление среднего на всем фрагменте речи приводит к большой задержке распознавания речи. Для того чтобы этого избежать, применяется фильтрация ИП, которая отличается от нормализации тем, что средние значения коэффициентов определяются 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 последующих фреймов. В ходе экспериментов была определена оптимальная длительность фрагмента усреднения — 5 с, обеспечивающая при минимальной задержке не менее 2,5 с наилучшее отношение повышения эффективности распознавания речи к минимальной задержке.

Как отмечалось в [4], вклад кепстральных коэффициентов со старшими значениями индексов в оценку меры близости между входным сигналом и моделью в системе распознавания речи невелик, а более значительная дисперсия первых кепстральных коэффициентов объясняется их большей зависимостью от частоты основного тона, определяемой типом эмоции [2].

Для устранения этих недостатков необходимо провести кепстральное взвешивание (L, см. рис. 1, — это количество кепстральных коэффициентов в ИП, $L = 12$).

Для возможности использования информации о динамике речи при верификации результатов распознавания в работе [9] вместе с исходными кепстральными коэффициентами в информативный признак введены параметры, характеризующие спектральные переходы. В качестве динамических параметров речи хорошо себя зарекомендовали первые и вторые производные кепстральных коэффициентов и логарифма энергии во фрейме [10].

Сформированный таким образом ИП можно считать устойчивым к эмоциям и инвариантным к диктору, а его размерность пространства признаков будет равна 39 (логарифм энергии во фрейме + 12 кепстральных коэффициентов + 13 первых + 13 вторых производных).

База эмоциональной речи и система распознавания. Для проведения эксперимента по распознаванию эмоциональной речи была собрана база русской эмоциональной слитной речи. База состоит из 13 текстов (продолжительностью от 20 до 50 секунд речи), которые записаны при участии 22 человек (16 мужчин и 6 женщин) в возрасте от 22 до 45 лет. Каждый текст записан с соответствующей эмоцией (гнев, радость, отвращение, удивление, печаль, страх и нейтральное эмоциональное состояние). Общий размер базы составляет 286 файлов или приблизительно 3 часа речевых данных. Запись всех файлов осуществлялась с частотой дискретизации сигнала 16 000 Гц, разрядностью квантования 16 бит и в формате звукового файла *Waveform Audio File Format* (WAV). База записывалась на конденсаторном микрофоне BEHRINGER C-2 (с частотным диапазоном 20—20 000 Гц и соотношением сигнал/шум 75 дБ) с использованием внешней звуковой карты Creative E-MU 0202 USB 2.0.

Обучение и тестирование проводилось на основе перекрестной проверки (*k-fold cross-validation* [11]) с разбиением речевых данных на десять равных частей. Эксперимент проводился по трем сценариям, согласно которым обучающая выборка состояла только из нейтральных или только из эмоциональных речевых данных, а также из нейтральных и эмоциональных речевых данных в соотношении 1:1.

Система распознавания речи реализована на основе скрытых марковских моделей. Значения параметров СММ оцениваются на обучающей выборке с ее ручной транскрипцией по фонемам. Результат распознавания получается путем выбора последовательности слов с максимальной апостериорной вероятностью.

Для определения эффективности распознавания эмоциональной речи была применена следующая формула:

$$W_{\text{acc}} = \frac{N - S - D - I}{N},$$

где N — число слов в распознаваемой речи (правильная транскрипция), S — число замененных слов в речи при распознавании, D — число удаленных слов из речи при распознавании, а I — число вставленных слов в речь при распознавании.

Сравнительный анализ результатов распознавания эмоциональной речи. Оценим эффективность описанного в статье ИП, полученного на основе кепстральных коэффициентов, определенных на экспоненциально-логарифмической шкале частот, для спектра, рассчитанного по параметрам линейного предсказания (ЛПСКК) в целом и относительное повышение эффективности распознавания эмоциональной речи при использовании 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 быстрым преобразованием Фурье (БПФ) и мел-частотной шкалой соответственно. В литературных источниках такой информативный признак называется мел-частотным кепстральным коэффициентом (МЧКК, *Mel-Frequency Cepstrum Coefficients* [5]).

В табл. 1 используются следующие обозначения: FxdP — коррекция фильтром первого порядка, CMFilt — фильтрация кепстральных коэффициентов, CepLift — кепстральное взвешивание, Δ — добавление первых и вторых производных. Результаты распознавания эмоциональной речи приводятся для трех случаев: Н — эффективность распознавания речи с нейтральным эмоциональным состоянием, Г — стрессовым состоянием (гнев) и О — для всех эмоциональных состояний, представленных в собранной базе (гнев, радость, отвращение, удивление, печаль, страх и нейтральное эмоциональное состояние).

Таблица 1

Информативный признак	ЛП-анализ			БПФ		
	Н, %	Г, %	О, %	Н, %	Г, %	О, %
	МЧКК					
FxdP	62,3	20,8	39,8	67,2	12,5	35,7
FxdP + CMFilt	65,8	27,1	45,9	70,8	15,4	41,3
FxdP + CMFilt + CepLift	66,7	29,2	47,6	71,3	16,1	42,9
FxdP + CMFilt + CepLift + Δ	70,9	32,5	52,2	75,7	19,8	**48,4**
	ЛПСКК					
FxdP	49,3	35,6	40,8	57,1	16,4	34,7
FxdP + CMFilt	55,8	42,1	49,7	58,8	20,6	39,9
FxdP + CMFilt + CepLift	57,1	42,9	50,4	60,9	21,1	40,3
FxdP + CMFilt + CepLift + Δ	61,6	47,8	**54,3**	64,4	23,7	45,1

Из табл. 1 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 позволит добиться повышения эффективности распознавания гласных звуков с различной эмоциональной окраской. С другой стороны, как и было сказано в [4], применение ЛП-анализа приводит к ошибочному определению 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 и добавления первых и вторых производных позволяет повысить эффективность распознавания эмоциональной речи. А сочетание ЛП-анализа с использованием экспоненциально-логарифмической шкалы частот (ЛПСКК) обеспечивает максимальную эффективность распознавания эмоциональной речи.

Рассмотрим для информативных признаков МЧКК и ЛПСКК влияние типа речевых данных обучающей выборки, на эффективность распознавания речи. Полученные результаты приведены в табл. 2 (Э — эффективность распознавания только для эмоциональных состояний, кроме нейтрального).

Таблица 2

Обучающая выборка	ЛПСКК		МЧКК	
	Н, %	Э, %	Н, %	Э, %
Н	63,7	48,5	**86,1**	14,6
Э	50,9	**57,2**	52,9	27,5
Н:Э (1:1)	61,6	52,3	75,7	25,4

Заключение. 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 5,9 % 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.

Эффективность системы распознавания эмоциональной речи может быть повышена путем добавления этапа предварительной сегментации тестовой выборки при помощи детектора эмоциональной речи. В этом случае, используя МЧКК (обученный на нейтральной речи) и ЛПСКК (обученный на эмоциональной речи) для декодирования нейтральной и эмоциональной речи соответственно, можно будет добиться максимальной эффективности распознавания речи.

СПИСОК ЛИТЕРАТУРЫ

1. *Vlasenko B., Prylipko D., Wendemuth A.* Towards robust spontaneous speech recognition with emotional speech adapted acoustic models // 35th German Conf. on Artificial Intelligence. German, 2012. P. 103—107.
2. *Williams C. E., Stevens K. N.* Emotions and speech: Some acoustical correlates // J. Acoust. Soc. Amer. 1972. N 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 // Speech and Audio Processing. 2000. N 8. P. 429—442.
4. *Рылов А. С.* Анализ речи в распознающих системах. Мн.: Бестпринт, 2003. 264 с.
5. *Huang X., Acero A., Hon H.-W.* Spoken language processing. New Jersey: Prentice-Hall, Inc., 2001. 980 p.
6. *Корн Г., Корн Т.* Справочник по математике для научных работников и инженеров. М.: Наука, 1974. 143 с.
7. *Chen Y.* Cepstral domain stress compensation for robust speech recognition // Proc. IEEE Intern. Conf. on Acoustics, Speech, and Signal Processing. Dallas, 1987. P. 717—720.
8. *Hansen J. H. L, Bria O. N.* Lombard effect compensation for robust automatic speech recognition in noise // Intern. Conf. Spoken Language Processing. Kobe, Japan, 1990. P. 1125—1128.
9. *Furui S.* Cepstral analysis technique for automatic speaker verification // IEEE Transact. on Acoustics, Speech and Signal Processing. 1981. Vol. 29, N 2. P. 254—272.
10. *Beulen K., Welling L., Ney H.* Experiments with linear feature extraction in speech recognition // Speech communication and technology: European Conf. Madrid, 1995. N 2. P. 1415—1418.
11. K-fold cross-validation [Электронный ресурс]: <http://en.wikipedia.org/wiki/Cross-validation_% 8statistics%29>.

Сведения об авторе

Андрей Владимирович Ткаченя — аспирант; Белорусский государственный университет, кафедра радиофизики и цифровых медиатехнологий;
E-mail: tkachenia@gmail.com

Рекомендована кафедрой радиофизики и цифровых медиатехнологий

Поступила в редакцию 02.10.14 г.

Ссылка для цитирования: *Ткаченя А. В.* Методика формирования устойчивых к эмоциям информативных признаков для задачи распознавания речи // Изв. вузов. Приборостроение. 2015. Т. 58, № 6. С. 443—450.

DEVELOPMENT OF EMOTION-TOLERANT INFORMATIVE INDICATORS FOR SPEECH RECOGNITION PROBLEM

A. V. Tkachenia

Belarus State University, 220030, Minsk, Republic of Belarus,
E-mail: tkachenia@gmail.com

A method of the speech signal parameterization providing emotion-tolerant and speaker-invariant feature vector is proposed. The method makes use of the cepstral coefficients defined on an ExpoLog frequency scale on the base of on a linear-prediction power spectrum. The described feature vector is applied for emotional speech recognition based on hidden Markov models. Experimental results demonstrate that the use of the proposed method improves emotional speech recognition efficiency by 5,9 %.

Keywords: emotional speech recognition, feature vector, linear prediction coefficients, cepstral coefficients, hidden Markov models.

Data on authors

Andrey V. Tkachenia — Post-Graduate Student; Belarus State University, Department of Radiophysics and Digital Media Technologies;
E-mail: tkachenia@gmail.com

Reference for citation: *Tkachenia A. V.* Development of emotion-tolerant informative indicators for speech recognition problem // Izvestiya Vysshikh Uchebnykh Zavedeniy. Priborostroenie. 2015. Vol. 58, N 6. P. 443—450 (in Russian).

DOI: 10.17586/0021-3454-2015-58-6-443-450