СИСТЕМА РАСПОЗНАВАНИЯ ЭМОЦИОНАЛЬНОЙ СЛИТНОЙ РЕЧИ НА ОСНОВЕ СКРЫТЫХ МАРКОВСКИХ МОДЕЛЕЙ

А. В. Ткаченя

Белорусский государственный университет
Минск, Беларусь
e-mail: tkachenia@gmail.com

Описываются разработанные методики (формирование инвариантного к эмоциям вектора признаков речевого сигнала, декодирование спонтанной речи при помощи триггерной сети спутывания, комбинированная верификация слов распознанной слитной речи и интерактивная неконтролируемая адаптация скрытых марковских моделей с механизмом обновления), совокупное использование которых обеспечивает увеличение точности предложенной системы распознавания эмоциональной слитной речи на основе скрытых марковских моделей. Полученные в ходе эксперимента результаты свидетельствуют, что рассматриваемая в статье система позволяет повысить точность распознавания эмоционально окрашенной речи в среднем на 14,5 %.

Ключевые слова: распознавание эмоциональной слитной речи; декодирование речи; верификация слов; адаптация скрытых марковских моделей.

EMOTIONAL CONTINUOUS SPEECH RECOGNITION SYSTEM BASED ON HIDDEN MARKOV MODELS

A. V. Tkachenia

Belarus State University
Minsk, Belarus

In the paper developed methods (emotional invariant feature vector parameterization, spontaneous speech decoding based on trigger confusion network, composite verifying the words of recognized continuous speech and on-line unsupervised adaptive learning of hidden Markov models with updating mechanism) are described, the joint usage of which provides an increase in the accuracy of the proposed emotional continuous speech recognition system based on hidden Markov models. Experimental results show that usage of the described system leads to improving the efficiency of emotional speech recognition by an average of 14.5 %.

Keywords: emotional continuous speech recognition; speech decoding; word verification; hidden Markov models adaptation.

ВВЕДЕНИЕ

В условиях бурного развития информационных технологий в современном мире постоянно растет актуальность проблемы быстрого взаимодействия человека и ком-

пьютера посредством речи. В связи с этим продолжают интенсивно развиваться и совершенствоваться методы распознавания речи, и одной из наиболее актуальных задач является разработка систем распознавания спонтанной речи, для которой характерно слитное и эмоциональное произношение. Приводимые в литературных источниках данные показывают [1], что в зависимости от сложности задачи и условий точность распознавания эмоциональной слитной речи лежит в пределах от 35 % до 65 %, что сильно уступает качеству распознавания нейтральной слитной речи, для которой указанная величина составляет 88 %.

Распознавание эмоциональной слитной речи – это сложная комплексная задача, для решения которой необходимо дать ответ на ряд актуальных проблем, которые можно разделить на три группы: акустическое моделирование, языковое моделирование и увеличение точности распознавания речи.

Для акустической теории речеобразования характерны два типа изменчивости: акустическая и временная. В увеличении точности распознавания эмоциональной речи центральную роль играет снижение акустической изменчивости характеристик речевого сигнала. Для этого используют инвариантные к эмоциям информативные признаки (ИП) [2, 3], методы компенсации эмоций в ИП [4, 5] и методы адаптации моделей [6]; два последних подхода требуют предварительной классификации эмоций. При этом средняя точность современных классификаторов эмоций составляет 59,9 % [7], что ставит под сомнение эффективность их использования для предварительного анализа речевого сигнала в системах распознавания эмоциональной речи. Ошибки, вносимые временной изменчивостью, можно снизить за счет использования скрытых марковских моделей (СММ) [8] и алгоритма динамической трансформации шкалы времени (ДТВ) [9], при помощи которых решается проблема различной длительности сигнала, соответствующего одной и той же фонеме.

Важности языкового моделирования эмоциональной речи и его влияния на точность распознавания посвящен ряд научных работ [10, 11]. Использование в качестве речевой единицы СММ трифонов со связанными состояниями позволяет хорошо моделировать слитную речь, избежать нехватки обучающих данных, а также решить проблему слов вне словаря. Для компактного представления результатов распознавания речи применяют сети спутывания [12], которые позволяют не только повысить эффективность декодирования речи, но и снизить вычислительные затраты. Определять контекстную связь между словами для задачи декодирования спонтанной речи следует на основе триггерной модели [13], главным достоинством которой является возможность автоматического обучения языковой модели.

ПОДХОДЫ К УВЕЛИЧЕНИЮ ТОЧНОСТИ РАСПОЗНАВАНИЯ РЕЧИ

В статье [14] была предложена методика формирования инвариантного к эмоциям вектора признаков при помощи линейного предсказания и экспоненциально-логарифмической шкалы частот. Использование спектра мощности речевого сигнала, рассчитанного на основе коэффициентов линейного предсказания, позволяет решить проблему изменения частоты основного тона в эмоциональной речи за счет оценки огибающей спектра, что обуславливает лучшее распознавание гласных с различной эмоциональной окраской. Однако отсутствие нулей в полученном спектре приводит к спутыванию схожих согласных, что сказывается на точности распознавания нейтральной речи. В ходе экспериментов также было установлено, что для эмоциональных состояний вторая форманта меньше всего изменяется по сравнению с нейтральным со-

стоянием. По этой причине для оценки кепстра линейного предсказания было предложено использовать экспоненциально-логарифмическую шкалу частот, которая позволяет снизить изменчивость пространства информативных признаков для эмоционально окрашенных речевых сигналов.

Для увеличения точности распознавания эмоциональной слитной речи была разработана методика декодирования спонтанной речи на основе сети спутывания и триггерной языковой модели [15]. В ней реализована генерация сети спутывания из графа слов на основе алгоритма выравнивания по опорным точкам Тура, который позволяет добиться оптимального соотношения между величиной ошибок выравнивания и вычислительными затратами. Как было показано в статье [15], предложенная методика обеспечивает эффективное автоматическое обучение триггерной языковой модели эмоциональной речи.

Повысить точность системы распознавания эмоциональной речи можно за счет верификации слов на основе алгоритма ДТВ. Уточненная апостериорная вероятность распознанного слова, определенная как линейная интерполяция апостериорной вероятности для триггерной сети спутывания P_{CN} и верификации на основе алгоритма ДТВ P_{DTW}, позволяет увеличить точность распознавания эмоциональной слитной речи [16]. Это обусловлено тем, что P_{CN}, полученная на основе СММ, хорошо описывает статистику акустических характеристик речи, тем самым являясь менее восприимчивой к их вариации в анализируемых данных, в то же время приводит к плохому распознаванию схожих слов. Напротив, P_{DTW} хорошо отражает изменения акустических характеристик в анализируемых данных, что дает возможность эффективно различать схожие слова, но порождает ошибки распознавания при вариации акустических характеристик речи.

Полученная на этапе верификации слов автоматическая транскрипция распознанной речи может быть использована при интерактивной неконтролируемой адаптации СММ, для увеличения эффективности которой был разработан механизм обновления [17] задающий степень доверия к новым входным данным, чтобы не допускать адаптацию СММ на ложно распознанных словах. Адаптация СММ позволяет снизить несоответствие между акустическими характеристиками в полученных моделях и тестовых данных, тем самым, повышая точность распознавания речи.

БАЗА РЕЧЕВЫХ СИГНАЛОВ И РЕЗУЛЬТАТЫ ЭКСПЕРИМЕНТОВ

Для оценки точности распознавания русской речи была собрана база эмоциональной слитной русской речи ЭМО-РУСС. База ЭМО-РУСС 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 текс записан в двух вариантах: в нейтральном эмоциональном состоянии и с требуемой эмоцией (гнев, страх, отвращение, печаль, удивление или радость). Общий размер базы составляет 572 файла. Запись всех файлов осуществлялась с частотой дискретизации сигнала 16000 Гц, разрядностью квантования 16 бит и в формате звукового файла *Waveform Audio File Format* (WAV). База записывалась на конденсаторном микрофоне BEHRINGER C-2 (частотный диапазон 20-20000 Гц и соотношением сигнал/шум 75 дБ) с использованием внешней звуковой карты Creative E-MU 0202 USB 2.0.

Используя описанные выше методики, повышающие точность распознавания эмоциональной слитной речи, была реализована система распознавания на основе СММ (EmotionalASR), процедурная блок-схема которой представлена на рис. 1.

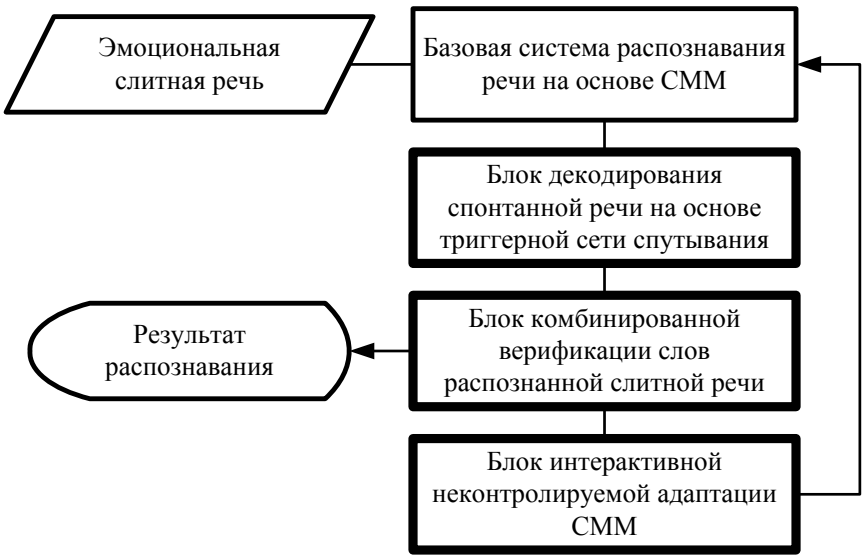

Рис. 1. Схематическая структура разработанной системы распознавания эмоциональной слитной речи на основе СММ

В ходе проведенных экспериментов (табл. 1) было показано, что интернет сервис Google ASR [18], коммерческий программный продукт Phonexia LVCSR [19] и библиотека функций с открытым исходным кодом CMU Sphinx [20] имеют более низкую точность распознавания эмоциональной слитной речи. В свою очередь, система Hybrid ANN/HMM ASR with DEA [21] показала неплохие результаты, что свидетельствует об эффективности использования гибридных алгоритмов распознавания речи, а также подхода адаптации моделей, на точность которого влияет надежность классификации эмоций в речевом сигнале. В целом эксперименты на русскоязычной базе ЭМО-РУСС показывают, что точность распознавания эмоциональной слитной речи для разработанной системы в среднем на 14,5 % выше.

Таблица 1

Точность распознавания эмоциональной слитной речи для различных систем

Система	Точность распознавания речи, %
EmotionalASR	74,6 ± 1,7
Hybrid ANN/HMM ASR with DEA	70,7 ± 1,3
Google ASR	57,2 ± 3,2
Phonexia LVCSR	59,8 ± 2,9
CMU Sphinx	52,6 ± 2,6

БИБЛИОГРАФИЧЕСКИЕ ССЫЛКИ

1. Hansen J. H. L., Patil S. Speech under stress: analysis, modeling and recognition // Speaker Classification : in 2 vol. / ed.: Ch. Müller. Berlin ; Heidelberg : Springer, 2007. Vol. 1 : Fundamentals, Features, and Methods. P. 108–137.

2. Gharavian D., Ahadi S. M. Recognition of emotional speech and speech emotion in farsi // Chinese Spoken Language Processing : Proceedings., the 5th International Symposium (ISCSLP 2006) : in 2 vol., Kent Ridge, Singapore, december 13–16, 2006. Kent Ridge, 2006. Vol. 2. P. 299–308.

3. Wöllmer M., Schuller B., Rigoll G. Feature frame stacking in RNN-based tandem ASR systems – learned vs. predefined context // International Speech Communication Association : Proceedings., the 12th Annual Conference (INTERSPEECH 2011), Florence, Italy, august 28–31, 2011. Florence, 2011. P. 1233–1236.

4. Clary G. J., Hansen J. H. L. Feature enhancement for multi-layer perceptron and semi-continuous hidden Markov model based classifiers using neural networks // Neural and Stochastic Methods in Image and Signal Processing : Proceedings of the SPIE. 1992. Vol. 1766. P. 529–540.

5. Hansen J. H. L. Morphological constrained enhancement with adaptive cepstral compensation (MCE-ACC) for speech recognition in noise and Lombard effect // Speech and Audio Processing. 1994. Vol. 2, iss. 4. P. 598–614.

6. Vlasenko B. V., Prylipko D., Wendemuth A. Towards robust spontaneous speech recognition with emotional speech adapted acoustic models // Artificial Intelligence : Proceedings., the 35th German Conference (KI-2012), Magdeburg, Germany, 2012. Magdeburg, 2012. P. 103–107.

7. Acoustic emotion recognition: a benchmark comparison of performances / B. Schuller [et al.] // Automatic Speech Recognition and Understanding : Proceedings., IEEE Workshop (ASRU 2009), Merano, november 13, 2009. Merano, 2009. P. 552–557.

8. Rabiner L. A tutorial on hidden Markov models and selected applications in speech recognition // Proceedings of the IEEE. 1989. Vol. 77, iss. 2. P. 257–286.

9. Rabiner L., Juang B.-H. Fundamentals of speech recognition. NJ ; Upper Saddle River : Prentice-Hall, Inc., 1993.

10. Athanaselis T., Bakamidis S., Dologlou I. Automatic recognition of emotionally coloured speech // World Academy of Science, Engineering and Technology. 2007. Vol. 1, № 12. P. 904–907.

11. ASR for emotional speech: clarifying the issues and enhancing performance / T. Athanaselisa [et al.] // Neural Networks: Emotion and Brain. 2005. Vol. 18, iss. 4. P. 437–444.

12. Wang H. L. Research on confusion network and side information for speech recognition : Ph. D. Thesis, Harbin Institute of Technology. China, 2007.

13. Chen Y., Chan K. P. Extended multi-word trigger pair language model using data mining technique // Systems, Man And Cybernetics : Proceedings., the 3rd IEEE International Conference (SMC'03) : in 2 vol. Washington, USA, October 5–8, 2003. Washington, 2003. Vol. 1. P. 262–267.

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

15. Ткаченя А. В. Декодирование речи на основе триггерной сети спутывания // Электроника Инфо. 2014. № 8 (110). С. 20–23.

16. Ткаченя А. В. Верификации результатов распознавания эмоциональной слитной речи // Электроника Инфо. 2014. № 7 (109). С. 32–34.

17. Ткаченя А. В. Адаптация скрытых марковских моделей к распознаванию эмоционально окрашенной речи // Информатика. 2014. № 3 (43). С. 21–27.

18. Google ASR [Electronic resource] : Google Web Speech API for Automatic Speech Recognition. Mode of access: https://google.com (date of access: 21.09.2014).

19. Phonexia LVCSR [Electronic resource] : API for Large Vocabulary Continuous Speech Recognition / Phonexia. Mode of access: https://www.phonexia.com (date of access: 22.12.2014).

20. CMU Sphinx [Electronic resource] : Open source speech recognition ToolKit / Carnegie Mellon University. Mode of access: http://www.cmusphinx.sourceforge.net (date of access: 15.01.2015).

21. Schuller B., Stadermann J., Rigoll G. Affect-robust speech recognition by dynamic emotional adaptation // Speech Prosody : Proceedings., the 3rd International Conference (SProSIG 2006), Dresden, Germany, may 2–5, 2006. Dresden, 2006. P. 455–459.