УДК 004.934.1

А.В. Ткаченя

АДАПТАЦИЯ СКРЫТЫХ МАРКОВСКИХ МОДЕЛЕЙ К РАСПОЗНАВАНИЮ ЭМОЦИОНАЛЬНО ОКРАШЕННОЙ РЕЧИ

Рассматривается алгоритм интерактивной неконтролируемой оценки параметров скрытых марковских моделей (СММ). Решается задача адаптации СММ к эмоционально окрашенной речи. Для увеличения достоверности уточненных параметров СММ предлагается механизм забывания и обновления. Приводятся функциональная блок-схема рассматриваемого алгоритма адаптации СММ, а также полученные результаты улучшения эффективности распознавания эмоциональной речи.

Введение

Скрытая марковская модель является стохастическим конечным автоматом, состоящим из конечного множества состояний $Q = \{q_1, q_2, \ldots, q_N\}$ и вероятностей непосредственных переходов между ними $A = \{a_{ij}\}$. Каждое такое состояние q_i связано с вектором признаков o_t при помощи матрицы вероятности наблюдения $B = \{b_i(o_t)\}$, каждый элемент которой является эмиссионной плотностью распределения вероятности $P(o_t|q_i)$. Также немаловажным параметром СММ является вероятность начального распределения состояний $\pi = \{\pi_1, \pi_2, \ldots, \pi_N\}$, которое задает вероятность того, что первым будет i-е состояние. Сумма π_i должна быть равна единице. Таким образом, оценив параметры $\lambda = (\pi, A, B)$, СММ можно использовать для моделирования последовательности векторов признаков $O = \{o_1, o_2, \ldots, o_T\}$ как кусочно-стационарного процесса, в котором каждый стационарный участок относится к определенному состоянию СММ. Такой подход обеспечивает моделирование динамической структуры речевой единицы (т. е. решает проблему различной длительности сигнала, соответствующего одной и той же фонеме).

Известно, что снижение эффективности распознавания речи связано с несоответствием акустических характеристик обучающих и тестируемых данных. При адаптации моделей на основе классических методов оценки параметров СММ, таких как максимизация вероятности [1–4] и максимизация апостериорной вероятности [5–7], обычно используются алгоритмы, которые требуют обработки всех доступных данных целиком. Тем не менее в ряде работ [6, 8, 9] было предложено использовать интерактивную адаптацию СММ, которая позволяет осуществлять последовательную адаптацию на данных, полученных из тестируемой выборки, на этапе распознавания. Этот подход исключает необходимость предварительного сбора данных для адаптации и позволяет осуществить уточнение параметров СММ для каждого конкретного случая.

Таким образом, с помощью данных из тестируемой выборки и их транскрипции, полученной в результате распознавания эмоциональной речи на основе СММ, может быть произведена адаптация СММ с гауссовым распределением значений матрицы вероятности наблюдения B на основе квазибайесовского обучения [6], алгоритм которого заключается в последовательном обновлении параметров скрытых марковских моделей [8]. Такой тип адаптации называется неконтролируемым в противовес контролируемой адаптации, при которой транскрипция всех данных заведомо известна. Так как полученная транскрипция не проверяется вручную, в ней могут содержаться ошибки распознавания, поэтому было предложено использовать механизм забывания [9] и обновления СММ.

Адаптация СММ позволяет снизить несоответствие между акустическими характеристиками полученных моделей и тестируемых данных, повысив, таким образом, эффективность распознавания речи. Возможность эффективного применения интерактивной неконтролируемой адаптации СММ базируется на следующих условиях:

1. Эффективность распознавания эмоциональной речи на исходных (неадаптированных) моделях должна быть достаточно высока, чтобы проводить корректное уточнение, так как адаптация неконтролируема (ручная проверка правильности распознавания и корректности транскрипции распознанных данных не проводится).

2. Вектор признаков должен меняться плавно, так как адаптация на резко меняющихся данных неэффективна.

3. Изменения акустических характеристик должны иметь макроструктуру, т. е. прослеживаться на большом количестве фреймов.

Эмоциональная речь характеризуется изменением пространства информативных признаков по сравнению с нейтральной речью в соответствии с выражаемой эмоцией. Эти изменения прослеживаются на длительном промежутке времени и носят квазистационарный характер, что позволяет считать их макрособытиями. Известно также [10], что кепстральные коэффициенты, которые вычисляются на основе спектральных коэффициентов, рассчитанных по параметрам линейного предсказания, наименее восприимчивы к изменению частоты основного тона и вследствие этого дают хорошую эффективность распознавания эмоциональной речи.

1. Алгоритм адаптации параметров СММ

Для того чтобы решить задачу уточнения параметров СММ, вначале необходимо задать вид распределения вероятности наблюдения $b_i(o_t)$. В большинстве случаев требуется моделировать параметры СММ, используя для вероятности наблюдений многомерное распределение с непрерывной плотностью. В то же время при решении некоторых задач можно работать и с последовательностями наблюдений, в которых распределение вероятностей наблюдения дискретно. Для решения задачи адаптации СММ к эмоционально окрашенной речи будем использовать непрерывную плотность распределения вероятности наблюдения.

Для большинства систем СММ с непрерывной плотностью обычно используется распределение вероятности наблюдения в виде гауссовой смеси. При этом каждый вектор признаков в момент времени t можно разбить на S независимых информативных потоков o_{st}, тогда формула для вычисления $b_i(o_t)$ будет иметь следующий вид:

$$b_i(o_t) = \prod_{s=1}^{S}\left[\sum_{m=1}^{M_i} c_{ism} N(o_{st}; \mu_{ism}, \Sigma_{ism})\right], \tag{1}$$

где M_s – число компонент гауссовой смеси в потоке s; c_{ism} – вес m-й компоненты; $N(o;\mu,\Sigma)$ – многомерное распределение Гаусса с вектором математического ожидания μ и ковариационной матрицей Σ:

$$N(o;\mu,\Sigma) = \frac{1}{\sqrt{(2\pi)^n |\Sigma|}} e^{-\frac{1}{2}(o-\mu)^T \Sigma^{-1}(o-\mu)}, \tag{2}$$

где n – размерность o.

Для оценки параметров СММ чаще всего используется алгоритм Баума – Велша, который, по сути, является алгоритмом максимизации среднего (*EM algorithm*) [11]. В свою очередь, алгоритм Баума – Велша может быть эффективно реализован при помощи так называемого *forward-backward*-алгоритма [12].

Тогда определим

$$\gamma_i(t) = P(q_t = i \mid O, \lambda) = \frac{\alpha_i(t)\beta_i(t)}{\sum_{j=1}^{N} \alpha_j(t)\beta_j(t)} \tag{3}$$

как вероятность нахождения в i-м состоянии в момент времени t для последовательности векторов признаков O и

$$\xi_{ij}(t) = P(q_t = i, q_{t+1} = j \mid O, \lambda) = \frac{\gamma_i(t) a_{ij} b_j(o_{t+1}) \beta_j(t+1)}{\beta_i(t)} \tag{4}$$

как вероятность нахождения в i-м состоянии в момент времени t и перехода в j-е состояние в момент времени $t + 1$, где α является *forward*-вероятностью, а β – *backward*-вероятностью.

Цель алгоритма Баума – Велша заключается в оценке новых параметров СММ на основании старых параметров СММ и вновь поступивших данных. Определим для гауссовой смеси вероятность *m*-й компоненты *i*-й смеси для независимого информативного потока *s* как

$$\gamma_{ism}(t) = \gamma_{is}(t)\frac{c_{ism}b_{ism}(o_{st})}{\sum_{m=1}^{M_i} b_{ism}(o_{st})} = P(q_t = i, X_{ist} = m \mid O, \lambda), \tag{5}$$

где X_{ist} – случайная величина, указывающая компоненту *i*-й смеси для независимого информативного потока *s* в момент времени *t*.

Тогда для гауссовой смеси правила обновления параметров СММ для случая *K* наблюдаемых последовательностей векторов признаков, где T_k – количество векторов признаков в *k*-й последовательности, могут быть определены следующим образом:

$$\hat{\pi}_i = \frac{\sum_{k=1}^{K}\gamma_{is}^k(1)}{K}; \tag{6}$$

$$\hat{c}_{ism} = \frac{\sum_{k=1}^{K}\sum_{t=1}^{T_k}\gamma_{ism}^k(t)}{\sum_{k=1}^{K}\sum_{t=1}^{T_k}\gamma_{is}^k(t)}; \tag{7}$$

$$\hat{\mu}_{ism} = \frac{\sum_{k=1}^{K}\sum_{t=1}^{T_k}\gamma_{ism}^k(t)o_{st}^k}{\sum_{k=1}^{K}\sum_{t=1}^{T_k}\gamma_{ism}^k(t)}; \tag{8}$$

$$\hat{\Sigma}_{ism} = \frac{\sum_{k=1}^{K}\sum_{t=1}^{T_k}\gamma_{ism}^k(t)\left(o_{st}^k - \mu_{ism}\right)\left(o_{st}^k - \mu_{ism}\right)^T}{\sum_{k=1}^{K}\sum_{t=1}^{T_k}\gamma_{ism}^k(t)}, \tag{9}$$

а a_{ij} рассчитывается на основании выражений (3) и (4) с учетом (1) как

$$\hat{a}_{ij} = \frac{\sum_{k=1}^{K}\sum_{t=1}^{T_k}\xi_{ij}^k(t)}{\sum_{k=1}^{K}\sum_{t=1}^{T_k}\gamma_i^k(t)}. \tag{10}$$

В данном случае для модифицированного алгоритма Баума – Велша, используемого для интерактивного неконтролируемого уточнения параметров СММ с механизмом забывания и обновления, можно записать следующую последовательность шагов:

1. Каждый параметр СММ, требующий уточнения, записать в форме, приведенной в выражениях (6) – (10).

2. С учетом выражения (1) посчитать *forward*- и *backward*-вероятности для всех состояний *i* и моментов времени *t*.

3. Для каждого состояния *i* и момента времени *t*, используя выражения (3) – (5) и вектор признаков, на котором происходит адаптация, оценить новые параметры СММ.

4. Вычислить апостериорную вероятность $P(O \mid \lambda^{n+1})$ с новыми параметрами СММ.

5. Прекратить процесс уточнения, если значение $\varphi P(O \mid \lambda^n) > \psi P(O \mid \lambda^{n+1})$, где φ и ψ – коэффициенты забывания и обновления соответственно, а $P(O \mid \lambda^n)$ – апостериорная вероятность, рассчитанная на старых параметрах СММ и вычисленная на втором шаге.

6. Сохранить новые параметры СММ и повторить шаги 1–6 с новыми значениями параметров СММ.

В результате будет получена адаптированная СММ, в которой будут уточнены матрица переходов (A), вероятность начального распределения состояний (π), а также вектор математических ожиданий и ковариационная матрица, соответствующие вероятности наблюдения (B) с учетом коэффициента забывания φ и обновления ψ, влияние которых на эффективность распознавания речи будет исследовано экспериментально.

2. Система распознавания эмоциональной речи с адаптацией СММ

Дополнив систему распознавания эмоциональной речи блоком адаптации СММ, можно добиться улучшения эффективности распознавания эмоциональной речи для последующих речевых данных, так как они уже будут анализироваться на уточненных СММ (рис. 2). Заметим, что для коротких речевых сигналов (до 6–10 слов) эффект адаптации СММ не будет наблюдаться из-за отсутствия достаточного количества данных, необходимых для уточнения параметров СММ.

Стоит отметить, что при быстрой смене эмоционального состояния все же происходят значительные изменения акустических характеристик сигнала. Это может приводить к тому, что все значения апостериорной вероятности распознанных слов во множестве спутывания будут иметь близкие величины. В результате происходит снижение эффективности распознавания эмоциональной речи. Для этого случая следует предусмотреть возможность возврата к исходной модели, полученной на обучающих данных, в которых все эмоции представлены в равной мере. Условием возврата к исходной модели может служить следующее неравенство:

$$\left(\max_{m=1,M_k} \left(P_k^m \right) - \sum_{m=1}^{M_k} P_k^m \Big/ M_k \right) < thr, \tag{11}$$

где M_k – количество слов в k-м множестве спутывания; P_k^m – апостериорная вероятность m-го распознанного слова в k-м множестве спутывания; thr – это порог, который определяется экспериментальным путем.

Механизм забывания характеризуется коэффициентом $0 < \varphi \le 1$, при $\varphi = 1$ механизм забывания исходных моделей не задействован. Коэффициент φ нужен для снижения эффекта влияния прошлых наблюдений относительно новых входных данных, чтобы учитывалась изменчивость параметров СММ. Аналогичные механизмы забывания были также предложены в работах [13, 14].

В свою очередь, механизм обновления характеризуется коэффициентом $0 \le \psi \le 1$, который задает степень доверия к новым входным данным, чтобы не допускать адаптацию СММ на ложно распознанных словах. Коэффициент ψ было предложено определять исходя из формулы

$$\psi = \begin{cases} 0 & \text{при} \quad P_k < 2BL - 1; \\ \dfrac{P_k - BL}{1 - BL} + 1 & \text{при} \quad 2BL - 1 \le P_k < BL; \\ 1 & \text{при} \quad P_k \ge BL, \end{cases} \tag{12}$$

графическое представление которой показано на рис. 1. Здесь BL – среднее значение апостериорной вероятности распознанного слова для корректно распознанных слов. Параметры BL, как и все остальные параметры, которые должны быть исследованы экспериментально, определяются на проверочных данных сразу после обучения системы и при анализе тестовой выборки не меняются.

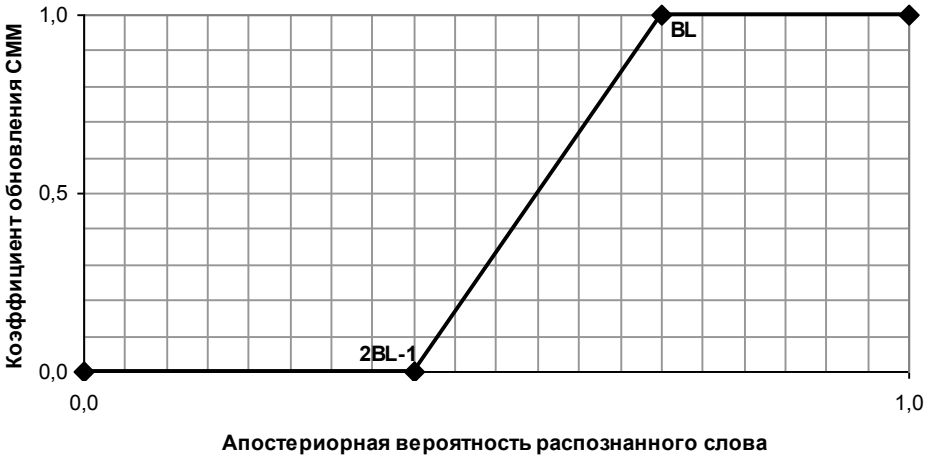

Рис. 1. Зависимость коэффициента обновления СММ от апостериорной вероятности распознанного слова

3. База эмоциональной речи и эффективность распознавания

Распознавание эмоциональной речи и оценка эффективности работы полученной системы будут осуществляться на японском просодическом корпусе MULTEXT (MULTEXT-J) [15]. Выбранная база состоит из 40 текстов (продолжительностью от 10 до 30 с речи), которые записаны при участии шести человек (трех мужчин и трех женщин) в возрасте от 20 до 50 лет. Каждый текст записан в двух вариантах: в нейтральном эмоциональном состоянии и с требуемой эмоцией. Таким образом, размер базы составляет 480 файлов. Запись всех файлов производилась с частотой дискретизации сигнала 16 000 Гц, разрядностью квантования 16 бит, в формате звуковых файлов Waveform Audio File Format (WAV).

Перед обучением и тестированием сигнал преобразовывался в последовательность векторов признаков, представляющих из себя 13 кепстральных коэффициентов, полученных на основе спектральных коэффициентов, рассчитанных по параметрам линейного предсказания. Эти кепстральные коэффициенты распределены по экспоненциально-логарифмической шкале частот, которая снижает изменчивость пространства информативных признаков для эмоциональной речи по сравнению с нейтральной [16]. В состав векторов были включены приближения первой и второй производной каждого коэффициента. Таким образом, общая размерность пространства векторов признаков составила 39.

Обучение и тестирование проводились на основе перекрестной проверки (k-fold cross-validation [17]) с разбиением базы на 10 равных частей. Исходные значения параметров СММ оцениваются на обучающей выборке с учетом ее ручной транскрипции по фонемам.

Для определения величины эффективности распознавания эмоциональной слитной речи была применена формула

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

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

Зависимость эффективности распознавания эмоциональной слитной речи от количества распознанных слов для разработанной системы (рис. 2) на базе MULTEXT-J представлена в таблице.

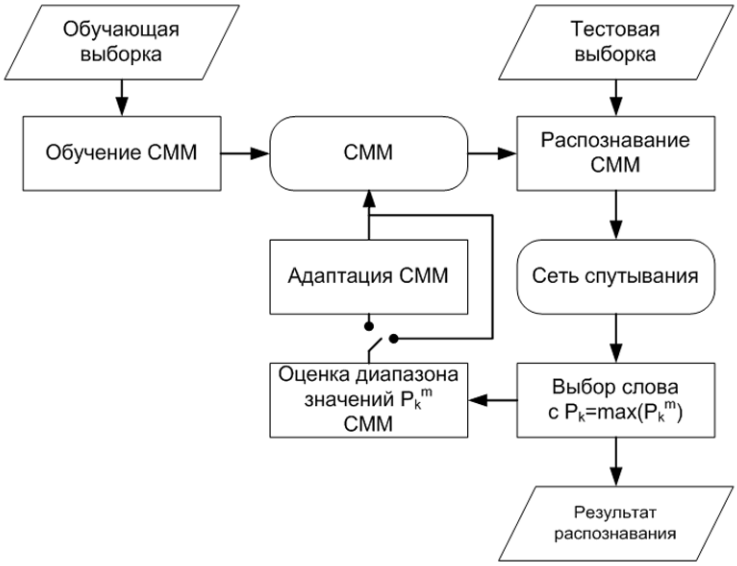

Рис. 2. Система распознавания эмоциональной речи с адаптацией СММ

Эффективность распознавания эмоциональной слитной речи в зависимости от количества распознанных слов

Количество распознанных слов, шт.	Эффективность распознавания речи, %
0	70
5	70
10	71
15	73
20	74
30	**75**
40	75
50	75

Отметим, что случай для нуля распознанных слов соответствует эффективности распознавания эмоциональной слитной речи при отсутствии адаптации СММ на данных из тестируемой выборки. Для предложенного вектора признака она равняется 70 %.

В ходе экспериментов была определена оптимальная величина коэффициента забывания φ, равная 0,75. Величина порога *thr* в формуле (11) для базы MULTEXT-J оказалась равной 0,1. Значение величины *BL* для определения коэффициента обновления ψ по формуле (12) составило 0,82.

Заключение

На основе изложенной в статье теории была создана система распознавания эмоциональной речи, дополненная блоком адаптации СММ с механизмом забывания и обновления. Как видно из таблицы, максимальная эффективность распознавания речи, равная 75 %, достигается при анализе 30 слов и более, что соответствует приблизительно 20 с эмоциональной речи.

Таким образом, предложенный дополнительный этап адаптации СММ с механизмом забывания и обновления позволил повысить эффективность распознавания эмоциональной речи на 5 %.

Список литературы

1. Baum, L.E. An inequality and associated maximization techniques in statistical estimation for probabilistic functions of Markov processes / L.E. Baum // Inequalities. – 1972. – № 3. – P. 1–8.

2. A maximization technique occurring in the statistical analysis of probabilistic function of Markov chains / L.E. Baum [et al.]// Ann. Math. Stat. – 1970. – № 41. – P. 164–171.

3. Juang, B.-H. Maximum likelihood estimation for multivariate mixture observations of Markov chains / B.-H. Juang, S.E. Levinson, M.M. Sondhi // IEEE Trans. Inform. Theory. – 1993. – № 2. – P. 307–309.

4. Liporace, L.R. Maximum likelihood estimation for multivariate observations of Markov sources / L.R. Liporace // IEEE Trans. Inform. Theory. – 1995. - № 28. – P. 729–734.

5. Gauvain, J.-L. Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains / J.-L. Gauvain, C.-H. Lee // IEEE Trans. Speech Audio Processing. – 1994. – № 2. – P. 291–298.

6. Huo, Q. Bayesian adaptive learning of the parameters of hidden Markov model for speech recognition / Q. Huo, C. Chan, C.-H. Lee // IEEE Trans. Speech Audio Processing. – 1992. – № 5. – P. 334–345.

7. Lee, C.-H. A study on speaker adaptation of the parameters of continuous density hidden Markov models / C.-H. Lee, C.-H. Lin, B.-H. Juang // IEEE Trans. Signal Processing. – 1991. – № 39. – P. 806–814.

8. Matsuoka, T. A study of on-line Bayesian adaptation for HMM-based speech recognition / T. Matsuoka, C.-H. Lee // Proc. EUROSPEECH-93. – Berlin, Germany, 1993. – P. 815–818.

9. Huo, Q. On-Line Adaptive Learning of the Continuous Density Hidden Markov Model Based on Approximate Recursive Bayes Estimate / Q. Huo, C.-H. Lee // Speech and Audio Processing. – 1997. – № 5. – P. 161–172.

10. Рылов, А.С. Анализ речи в распознающих системах / А.С. Рылов. – Минск : Бестпринт, 2003. – 264 с.

11. Bilmes, J. A Gentle Tutorial of the EM Algorithm and its Application to Parameter Estimation for Gaussian Mixture and Hidden Markov Models / J. Bilmes // International Computer Science Institute. – 1998. – № 1. – P. 164–191.

12. The HTK Book (for HTK v. 3.4) / S. Young [et al.]. – Cambridge University Engineering Department, 2006. – 359 p.

13. Krishnamurthy, V. On-line estimation of hidden Markov model parameters based on the Kullback-Leibler information measure / V. Krishnamurthy, J.B. Moore // IEEE Trans. Signal Processing. – 1993. – № 41 (8). – P. 2557–2573.

14. Weinstein, E. Sequential algorithms for parameter estimation based on the Kullback-Leibler information measure / E. Weinstein, M. Feder, A.V. Oppenheim // IEEE Trans. Acoust, Speech, Signal Processing. – 1990. – № 38 (9). – P. 1652–1654.

15. MULTEXT-J. Japanese MULTEXT Prosodic Corpus [Electronic resource]. – Mode of access : http://research.nii.ac.jp/src/en/MULTEXT-J.html. – Date of access : 30.09.2013.

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

17. K-fold cross-validation. Wikipedia [Electronic resource]. – Mode of access : http://en.wikipedia.org/wiki/Cross-validation_%28statistics%29. – Date of access : 18.08.2014.

Поступила 16.06.2014

Белорусский государственный университет,
Минск, ул. Курчатова, 5
e-mail: tkachenia@gmail.com

A.V. Tkachenia

ADAPTIVE LEARNING OF HIDDEN MARKOV MODELS FOR EMOTIONAL SPEECH

An on-line unsupervised algorithm for estimating the hidden Markov models (HMM) parameters is presented. The problem of hidden Markov models adaptation to emotional speech is solved. To increase the reliability of estimated HMM parameters, a mechanism of forgetting and updating is proposed. A functional block diagram of the hidden Markov models adaptation algorithm is also provided with obtained results, which improve the efficiency of emotional speech recognition.