УДК 519.68:007.5

Моделирование речевых признаков с помощью алгоритма симуляции отжига

А. В. Ермилов

Национальный исследовательский университет «Высшая школа экономики»
Кафедра управления разработкой программного обеспечения
ул. Мясницкая, д. 20, Москва, Россия, 101000

Мел-частотные кепстральные коэффициенты до сих пор являются наиболее популярными речевыми признаками. Однако в зависимости от длины речевого тракта (стоит отметить, что длина речевого тракта зависит от пола и других физиологических параметров, таких как рост, и может меняться в пределах от 13 до 18 см) частоты центральных формант оказываются смещёнными. Величина смещения может достигать 25%. Такие большие различия могут вести к неправильному распознаванию высказывания предварительно хорошо обученной модели в случае, если высказывание было произнесено новым диктором, то есть система становится дикторозависимой. Альтернативой является применение признаков, которые не зависят от диктора, например, полученные с помощью аудиовизуальных моделей (Auditory Image Model).

В данной статье описываются признаки, основанные на аудиовизуальных моделях, которые могут быть вычислены при помощи алгоритма симуляции отжига. На основе Монте-Карло-симуляций исследованы статистические свойства оценок параметров расширения Грам-Шарлье нормального распределения, полученных применением метода симуляции отжига к решению задачи максимизации правдоподобия, а также проведено сравнение точности решения данной задачи максимизации правдоподобия при помощи различных методов.

Ключевые слова: речевые признаки, алгоритм симуляции отжига, распознавание речи, моделирование распределений, численные методы.

1. Введение

Наиболее часто в системах распознавания речи используются мел-частотные кепстральные коэффициенты [1]. Однако из-за различий в длинах речевого тракта может происходит сдвиг частот центральных формант. Разница в этих частотах может доходить до 25%. Из-за этого различия первоначально обученная модель может плохо распознавать сообщения нового диктора, то есть система становится дикторозависимой.

Одним из способов решения этой проблемы является использование признаков, которые не меняются от диктора к диктору. В качестве таких признаков можно использовать признаки из Auditory Image Model (AIM) (см. описание в [2]).

В данной статье исследуется практическая применимость алгоритма симуляции отжига для отыскания оценок максимального правдоподобия параметров расширения Грам–Шарлье нормального распределения, используемого для моделирования речевых признаков, полученных по AIM.

2. Расширение Грам–Шарлье

Если истинная плотность распределения случайной величины z неизвестна, то разумно представить её в виде:

$$g(z) = p_n(z)\psi(z),$$

где $\psi(z)$ — плотность стандартного нормального распределения, а $p_n(z)$ выбрана таким образом, чтобы $g(z)$ имела те же моменты, что и истинная плотность z. Такая аппроксимация носит название расширения Грам–Шарлье.

Статья поступила в редакцию 27 сентября 2013 г.

Полиномы Эрмита образуют ортогональный базис относительно скалярного произведения, порождённого математическим ожиданием, взятым по плотности стандартного нормального распределения. Это свойство позволяет использовать многочлены Эрмита H_i в функции $p_n(z)$. К сожалению, полученная функция не является в строгом смысле плотностью: для некоторых значений параметров функция может принимать отрицательные значения. Для решения этой проблемы в [3] предложено использовать положительную плотность:

$$g(z) = \psi(z)\frac{\left(1 + \sum\limits_{i=1}^{n} c_i H_i(z)\right)^2}{k},$$

где $k = 1 + \sum\limits_{i=1}^{n} c_i^2 i!$.

Подобная плотность удобна не только с теоретической точки зрения, но и с практической — при оценке параметров методом максимального правдоподобия логарифмическая функция правдоподобия получается разделяемой и содержит логарифмы положительных выражений, что упрощает численную оптимизацию:

$$\ell = \ln(\psi(z)) + \ln\left(1 + \sum_{i=1}^{n} c_i H_i(z)\right)^2 - \ln\left(1 + \sum_{i=1}^{n} c_i^2 i!\right). \tag{1}$$

3. Монте-Карло-эксперименты

Для анализа применимости алгоритма симуляции отжига для решения задачи нахождения оценок максимального правдоподобия плотности расширения Грам–Шарлье предлагается использовать метод Монте-Карло.

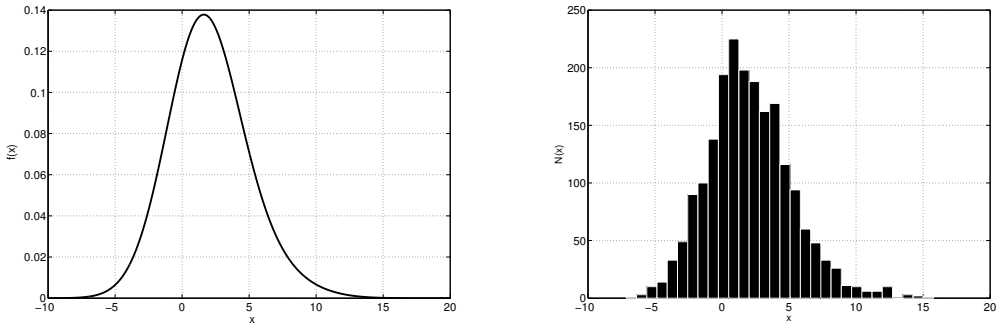
В качестве исследуемой плотности возьмём расширение Грам–Шарлье нормальной плотности с кумулянтами $\kappa_1 = 2, \kappa_2 = 3, \kappa_3 = 6, \kappa_4 = 10$. Для того, чтобы получить выборку из распределения с данной плотностью, был использован метод Монте-Карло по схеме марковской цепи (Monte-Carlo Markov's Chain, MCMC), при этом выборку сгенерируем по алгоритму срезов [4]. В ходе симуляций были исключены из выборки первые 5000 наблюдений для того, чтобы марковская цепь сошлась к своему стационарному распределению. Для устранения автокорреляций в выборке, сгенерированной марковской цепью, в итоговой выборке было оставлена каждое 5 сгенерированное значение. Всего было сгенерировано 2000 наблюдений.

График плотности изображён на рис. 1а. На рис. 1б представлена гистограмма наблюдений из распределения с указанной плотностью.

Для отыскания оценок методом максимального правдоподобия необходимо найти максимум функции (1). В нашем случае эта функция четырёх переменных.

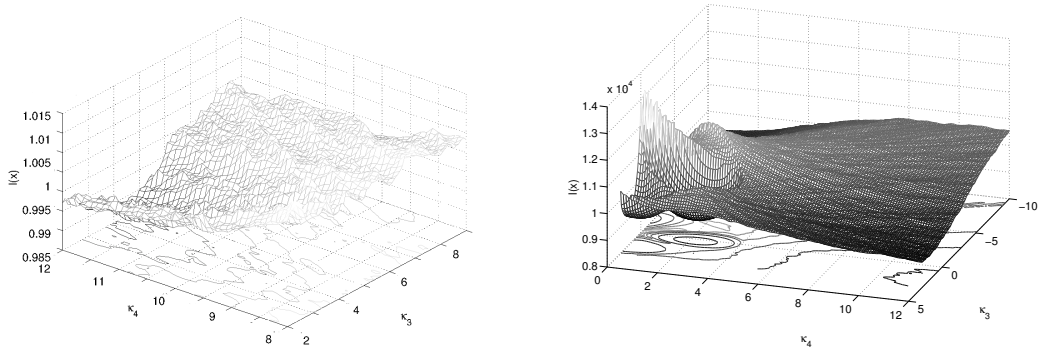
Особый интерес представляют оценки кумулянтов κ_3 и κ_4, описывающие асимметрию и эксцесс распределения. Из представленных графиков видно, что в случае некорректных значений параметров κ_1 и κ_2 (рис. 2а) целевая функция имеет очень много локальных минимумов и не является ни гладкой, ни выпуклой книзу. При этом даже при верных значениях κ_1 и κ_2 (рис. 2б) функция имеет локальные экстремумы.

В табл. 1 приведены оценки, полученные с помощью методов градиентного спуска, Нелдера–Мида [5] и симуляции отжига (при ограничении в 5000 вычислений значения функции). Из таблицы видно, что классические алгоритмы достаточно плохо справляются с оцениванием κ_4: оценки получились незначимыми на 95% уровня. Следует отметить, что в реальных ситуациях, когда наблюдения будут искажены шумами, целевая функция может иметь гораздо более сложную структуру, и результаты оценивания могут быть ещё более ненадёжными.



а б

Рис. 1. График плотности расширения Грам–Шарлье нормального распределения с кумулянтами $\kappa_1 = 2, \kappa_2 = 3, \kappa_3 = 6, \kappa_4 = 10$ (а) и выборка из этого распределения (б)



а б

Рис. 2. Вид отрицательной функции правдоподобия при фиксированных $\kappa_1 = 0$, $\kappa_2 = 1$ (а) и $\kappa_1 = 2$, $\kappa_2 = 3$ (б)

Для оценивания работы алгоритма построены эмпирические функции распределения оценок параметров, полученных методом симуляции отжига при числе перезапусков $N = 1000$ (рис. 3). Из полученных результатов видно, что получить надёжную оценку параметра κ_4 не получается даже при большом числе повторений.

Таблица 1

Оценки параметров, полученные разными методами

Параметр	Метод градиентного спуска	Метод Нелдера-Мида	Метод симуляции отжига
κ_1	2.04 (0.07)	2.02 (0.07)	1.97 (0.07)
κ_2	3.01 (0.05)	3.01 (0.05)	2.94 (0.05)
κ_3	5.4 (0.84)	5.38 (0.85)	5.35 (0.84)
κ_4	3.82 (5.1)	6.03 (5.12)	9.65 (5.84)

4. Заключение

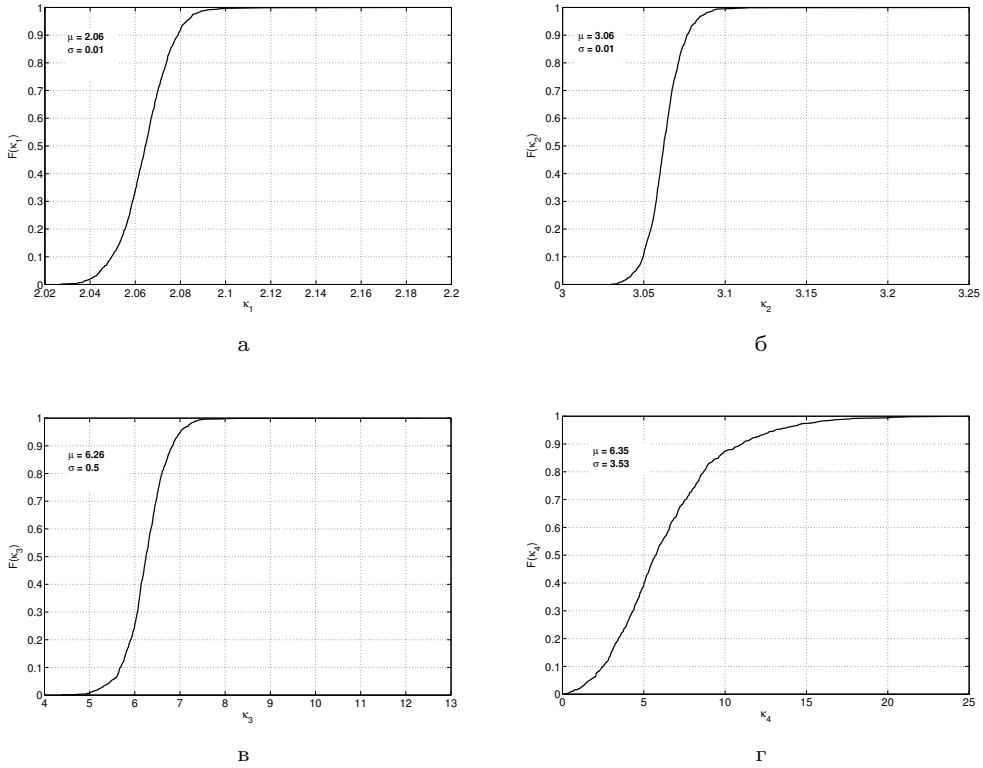


Рис. 3. Эмпирические функции распределения оценок параметров, полученных методом симуляции отжига

В данной статье исследована возможность оценивания параметров распределения профиля NAP, полученного по AIM-модели. Параметры данного распределения могут использоваться в системах автоматического распознавания речи как дикторонезависимые признаки. Мы построили и проанализировали эмпирические распределения оценок параметров, полученных с помощью метода симуляции отжига. Также проведено сравнение оценки, полученной методом симуляции отжига, с оценками, полученными классическими градиентными (метод градиентного спуска) и безградиентными методами (метод Нелдера–Мида).

В качестве дальнейшей работы можно предложить исследование возможности применения методов симуляций по схеме марковской цепи к оцениванию параметров смесей, модификаций EM-алгоритма для оценивания смеси распределений из расширений Грам–Шарлье нормального распределения, прямого применения вычислительных эвристик для моделирования смесей, а также сравнения работы этих алгоритмов.

Кроме того, интересно сравнить системы автоматического распознавания речи, построенных на смесях нормального и расширений нормального распределения, по скорости работы и точности распознавания.

Литература

1. *Sahidullah M., Saha G.* Design, Analysis and Experimental Evaluation of Block Based Transformation in MFCC Computation for Speaker Recognition // Speech Communication. — 2012. — Vol. 54, No 4. — Pp. 543–565.

2. *Munich M. E., Lin Q.* Auditory Image Model features for Automatic Speech Recognition // 9th European Conference on Speech Communication and Technology (Interspeech'2005 — Eurospeech). — 2005. — Pp. 3037–3040.
3. *Niguez T., Perote J.* Forecasting the Density of Asset Returns // STICERD Working Paper. — 2004.
4. *Neal R. M.* Slice Sampling // Annals of Statistics. — 2003. — Vol. 31, No 3. — Pp. 705–767.
5. Convergence Properties of the Nelder-MeadSimplex Method in Low Dimensions / J. C. Lagarias, J. A. Reeds, M. H. Wright, P. E. Wright // SIAM Journal on Optimization. — 1998. — Vol. 9, No 1. — Pp. 112–147.

UDC 519.68:007.5

Modeling Speach Features Via Simulated Annealing Algorithm

A. V. Ermilov

National Research University "Higher School of Economics"
Department of Control of System Development
20, Myasnickaya str, Moscow, Russia, 101000

Mel-Frequency Cepstral Coefficients are in so far the most popular speech features. However, depending on the length of a vocal tract (it is worth mentioning that length of a vocal tract is dependent on sex and other physiologic parameters of a speaker, such as height, and can vary from 13 cm to 18 cm) frequencies of central formants are shifted. The value of the shift can be as large as 25%. This huge difference can lead to a wrong recognition of a new utterance by a previously well-trained model when the utterance was said by a new speaker, thus the system becomes speaker-dependent. Alternative way is to use speaker independent features such as that obtained using Auditory Image Model (AIM) to describe input utterance.

In our work we propose AIM based features which are calculated using simulated annealing algorithm. Using Monte-Carlo schemes we investigate statistical properties of maximum likelihood estimates of Gram-Charlier extension of normal density obtained via simulated annealing algorithm, also we compare different methods to solve aforementioned optimization problem.

Key words and phrases: speach features, simulated annealing, speech recognition, distribution modeling, numerical methods.