УДК 004.93+57.087.1

АГЛОМЕРАТИВНАЯ КЛАСТЕРИЗАЦИЯ РЕЧЕВЫХ СЕГМЕНТОВ ФОНОГРАММЫ НА ОСНОВЕ БАЙЕСОВСКОГО ИНФОРМАЦИОННОГО КРИТЕРИЯ

О.Ю. Кудашев

Дано описание реализации системы агломеративной кластеризации речевых сегментов фонограммы на основе байесовского информационного критерия. Приведены результаты численных экспериментов с применением различных акустических признаков, а также с использованием полной и диагональной матриц ковариации. Для аудиозаписей радио «Свобода» на разработанной системе был достигнут уровень ошибки DER 6,4%.

Ключевые слова: кластеризация речевых сегментов, вариационный байесовский анализ, речевые технологии.

Введение

В последнее время наблюдается значительный рост интереса к системам автоматической сегментации фонограмм. Подобный рост вызван, в первую очередь, значительным увеличением объема звуковых данных, а также быстрым развитием технологий обработки речи. В значительной степени интерес научного сообщества в этой области поддерживается Национальным институтом стандартов и технологий (National Institute of Standards and Technology, NIST), разработавшим методику оценки качества систем автоматической обработки речи (Rich Transcription Evaluation Project, RTE) [1]. Одной из подзадач RTE является задача разделения дикторов (MDE (Metadata Extraction) Speaker Diarization), в рамках которой необходимо произвести кластеризацию (объединение) речевых сегментов фонограммы, принадлежащих одному диктору.

Задачи разделения дикторов и методы их решения можно классифицировать в соответствии с областью их применения. Так, для кластеризации речевых сегментов аудиозаписей диалогов наиболее эффективным методом является вариационный байесовский анализ в пространстве собственных голосов [2, 3]. Интересом данной работы являются аудиозаписи радиовещаний. Особенностью таких аудиозаписей является относительно редкая смена дикторов, а также отсутствие ограничения на их количество. Общепризнанным решением в этом случае является метод агломеративной кластеризации речевых сегментов на основе байесовского информационного критерия (Bayesian Information Criterion, BIC). Системы разделения дикторов, основанные на этом методе, зарекомендовали себя с точки зрения оптимального соотношения эффективности и производительности [4, 5].

Целью настоящей работы является реализация системы агломеративной кластеризации речевых сегментов на основе BIC. Кроме того, в данной работе будут представлены результаты применения разработанной системы к русскоязычному корпусу, в частности, к аудиозаписям радио «Свобода».

Применение BIC для кластеризации речевых сегментов

BIC является широко распространенным статистическим критерием, на основе которого производится выбор модели. В соответствии с этим критерием качество модели M, описывающей распределение данных $X=\{x_1, \ldots, x_N\}$, $x_i \in R^d$, оценивается по формуле

$$BIC(M) = \log L(x_1,...,x_N \mid M) - \frac{\lambda}{2}\nu(M)\log N, \quad (1)$$

где $L(x_1,...,x_N \mid M)$ – функция правдоподобия; $\nu(M)$ – число степеней свободы модели M (число свободных параметров); λ – настраиваемое пороговое значение, теоретически равное 1.

Пусть даны два независимых набора данных X_1, X_2. Задача кластеризации наборов данных X_1, X_2 может быть сведена к задаче выбора модели. Рассмотрим две альтернативные модели:

1. модель M_1 – данные X_1 и X_2 подчиняются одному гауссову распределению, $\{X_1, X_2\} \sim N(\mu, \Sigma)$.
2. модель M_2 – данные X_1 и X_2 подчиняются двум различным гауссовым распределениям, $\{X_1\} \sim N(\mu_1, \Sigma_1)$, $\{X_2\} \sim N(\mu_2, \Sigma_2)$.

Тогда, в соответствии с формулой (1),

$$BIC(M_1) = \log L(X_1, X_2 \mid \mu, \Sigma) - \frac{\lambda}{2}\nu(M_1)\log(N_1 + N_2),$$

$$BIC(M_2) = \log L(X_1 \mid \mu_1, \Sigma_1) + \log L(X_2 \mid \mu_2, \Sigma_2) - \frac{\lambda}{2}\nu(M_2)\log(N_1 + N_2),$$

где N_1, N_2 – количество данных в наборах X_1 и X_2 соответственно.

Разницу между этими двумя значениями обозначим ΔBIC:

$$\Delta BIC = BIC(M_1) - BIC(M_2) = \log\frac{L(X_1, X_2 \mid \mu, \Sigma)}{L(X_1 \mid \mu_1, \Sigma_1)L(X_2 \mid \mu_2, \Sigma_2)} + \frac{\lambda}{2}\left(\nu(M_2) - \nu(M_1)\right)\log N. \quad (2)$$

Положительное значение величины ΔBIC свидетельствует о том, что модель M_1 является наиболее предпочтительной, следовательно, наборы данных X_1 и X_2 следует отнести к одному кластеру.

Формулу (2) можно переписать в следующем виде:

$$\Delta BIC = \frac{1}{2}\left(N_1 \log\left|\Sigma_1\right| + N_2 \log\left|\Sigma_2\right| - (N_1 + N_2)\log\left|\Sigma_{1,2}\right|\right) + \frac{\lambda}{2}\beta\log(N_1 + N_2), \quad (3)$$

где Σ_1, Σ_2 – ковариационные матрицы данных X_1 и X_2 соответственно; $\Sigma_{1,2}$ – ковариационная матрица объединенных данных $\{X_1, X_2\}$; β – число свободных параметров, для диагональной ковариационной матрицы $\beta = d$, для полной ковариационной матрицы $\beta = \frac{d(d+1)}{2}$.

Алгоритм агломеративной кластеризации речевых сегментов

Алгоритм агломеративной кластеризации заключается в последовательном иерархическом попарном объединении (агломерации) кластеров (наборов речевых сегментов), принадлежащих одному диктору. При этом в качестве начальных кластеров используются речевые сегменты фонограммы, ограниченные двумя точками смены дикторов. Как для поиска точек смены дикторов, так и для определения принадлежности двух кластеров одному диктору используется BIC.

Разработанный алгоритм агломеративной кластеризации состоит из 4 этапов.

1. Выделение речевых сегментов фонограммы.
2. Расчет акустических признаков по всей фонограмме.
3. Поиск точек смены дикторов.
    1. Акустические признаки, лежащие внутри выделенных речевых сегментов, последовательно объединяются в непрерывный массив данных. В дальнейшем работа идет только в рамках этого массива.
    2. По всему полученному массиву перемещается окно фиксированной длины $2w$ с фиксированным шагом h. Окно разбивается на две равные части, левой части окна соответствуют данные X_1, правой – X_2 (формула (3)). Для точки $h \cdot i + w$, являющейся серединой окна, рассчитывается величина ΔBIC_i при $\lambda = \lambda_{CPD}$.
    3. Среди всех полученных таким образом значений выбираются локальные минимумы $M = \{m : \Delta BIC_{m-1} > \Delta BIC_m < \Delta BIC_{m+1}, \Delta BIC_m < 0\}$.

4. Среди всех полученных локальных минимумов M выбирается значение $m' = \arg\min_{m \in M}(\Delta BIC_m)$. При этом локальные минимумы из множества M, располагающиеся к выбранному значению ближе, чем w, удаляются: $M = M \setminus \{m : |m - m'| < w\}$.
5. Пункт 4 повторяется до тех пор, пока множество M не пусто.
6. Точки $\{h \cdot m' + w\}$ берутся в качестве точек смены дикторов.

4. Агломеративная кластеризация.
   1. Формируется набор кластеров $C = \{c_k\}$, каждый элемент которого состоит из множества речевых сегментов, ограниченных двумя соседними точками смены дикторов.
   2. Подсчитываются попарные значения $\{\Delta BIC_{k,l}\}_{k,l=1}^{|C|}$ по формуле (3) между всеми кластерами при $\lambda = \lambda_{AC}$.
   3. Если существует пара кластеров $c_{k*}, c_{l*} : \{k*, l*\} = \arg\max_{k,l}(\Delta BIC_{k,l}), \Delta BIC_{k*,l*} > 0$, то они объединяются в один кластер $c_{k*} = c_{k*} \bigcup c_{l*}$, при этом кластер c_{l*} удаляется: $C = C \setminus \{c_{l*}\}$.
   4. Пункты 2–3 повторяются до тех пор, пока происходят объединения.
   5. Получившийся набор кластеров C и будет являться решением задачи разделения дикторов.

Важно отметить необходимость использования различных пороговых значений при подсчете величины ΔBIC на этапе определения точек смены дикторов (λ_{CPD}) и на этапе агломеративной кластеризации (λ_{AC}). Эта необходимость обусловлена тем, что при поиске точек смены дикторов ключевым требованием является низкий уровень ошибки пропуска. Возникающий в этом случае высокий уровень ошибки ложного срабатывания компенсируется на этапе агломеративной кластеризации.

Результаты численных экспериментов

В качестве тестовой базы использовались 20 60-минутных аудиозаписей радио «Свобода» [6]. Все аудиозаписи имеют один канал с частотой дискретизации 16000 Гц. Для всей базы тестирования была создана эталонная разметка с указанием имен дикторов и принадлежащих им речевых сегментов. Речевые сегменты каждого диктора были дополнительно разделены на 4 категории:

1. чистая речь (71,86%);
2. речь на фоне шума (8,66%);
3. речь на фоне речи (11,20%);
4. речь на фоне музыки (8,28%).

Среднее количество дикторов в одном файле базы тестирования составило 25.

Для численных экспериментов речевые сегменты брались из эталонной разметки без использования какой-либо дополнительной обработки.

В качестве акустических признаков использовались:

- мел-частотный банк фильтров (MBF);
- мел-частотные кепстральные коэффициенты (MFCC).

Для построения акустических признаков были использованы следующие характеристики:

- окно быстрого преобразования Фурье (БПФ) – 16 мс;
- шаг окна БПФ – 10 мс;
- частотный диапазон для банка фильтров – от 0 до 8000 Гц.

Дополнительно к акустическим признакам были добавлены логарифм энергии (E), производные первого (Δ) и второго (Δ^2) порядков.

Для оценки качества системы использовался стандартный показатель вероятности ошибки разделения дикторов (Diarization Error Rate, DER), используемый в NIST RTE 2006 [1]:

$$DER = \frac{\sum_S \left\{ dur(S) \cdot \left(\max\left[N_{ref}(S), N_{sys}(S) \right] - N_{correct}(S) \right) \right\}}{\sum_S \left\{ dur(S) \cdot N_{ref}(S) \right\}} \cdot 100\%,$$

где S – непрерывный речевой сегмент фонограммы; $dur(S)$ – длина этого сегмента; $N_{ref}(S)$ – число дикторов, которым принадлежит речевой сегмент S, в соответствии с эталонной разметкой; $N_{sys}(S)$ – число дикторов, которым принадлежит речевой сегмент S, в соответствии с полученной разметкой; $N_{correct}(S)$ – число корректно определенных дикторов на речевом сегменте S.

Размер окна w для поиска точек смены дикторов брался равным 3 с, шаг h брался равным 0,2 с.

Результаты численных экспериментов, представляющих зависимость величины DER от используемых акустических признаков и типа ковариационной матрицы, представлены в табл. 1.

Акустические признаки	Размерность признаков	Полная ковариационная матрица			Диагональная ковариационная матрица		
		λ_{CPD}	λ_{AC}	DER (%)	λ_{CPD}	λ_{AC}	DER (%)
MBF	20	1,2	3,2	**8,0**	5	30	21,1
MFCC	20	1,0	3,5	**7,7**	4,8	27	8,5
MFCC + E	20	1,0	3,3	**7,8**	4,8	27	8,3
MFCC + E + Δ	40	0,6	1,3	7,3	3	18	**6,4**
MFCC + E + Δ + Δ^2	60	0,55	1,1	7,2	2,3	14	**6,6**

Таблица 1. Зависимость DER от акустических признаков и типа ковариационной матрицы

Результаты численных экспериментов, представляющих зависимость относительной производительности системы от типа ковариационной матрицы и размера акустических признаков, представлены в табл. 2, где относительная производительность измеряется как отношение продолжительности всех речевых сегментов фонограммы ко времени работы системы. Все измерения были проведены при работе алгоритма на одном ядре процессора Intel Core i5 760 2.8 GHz.

Размерность акустических признаков	Относительная производительность	
	Полная ковариационная матрица	Диагональная ковариационная матрица
20	210	**3600**
40	56	**1900**
60	23,5	**1280**

Таблица 2. Зависимость относительной производительности системы от типа ковариационной матрицы и размера акустических признаков

Заключение

Разработанная система агломеративной кластеризации речевых сегментов фонограммы представляет собой эффективное решение с точки зрения соотношения эффективности и производительности. Как показывают численные эксперименты, данная система способна показать значение ошибки DER равной 6,4% на русскоязычных аудиозаписях радиовещания, где происходит редкая смена дикторов. Как показано в табл. 2, относительная производительность системы сильно зависит от применения полной или диагональной ковариационной матрицы. Интересным представляется тот факт, что в случае использования слабо коррелирующих акустических признаков мел-частотных кепстральных коэффициентов применение диагональной ковариационной матрицы способно привести к уменьшению ошибки. При использовании полной ковариационной матрицы увеличение размерности акустических признаков приводит к сильному падению производительности прямо пропорционально квадрату размерности признаков, что еще раз демонстрирует целесообразность использования диагональной ковариационной матрицы.

Данная система разработана на кафедре «Речевые информационные системы», являющейся базовой кафедрой компании ООО «Центр речевых технологий». Она успешно применяется в системах автоматической обработки и распознавания речи.

Литература

1. Rich Transcription Evaluation Project [Электронный ресурс]. – URL: http://www.itl.nist.gov/iad/mig/tests/rt/, свободный. Яз. англ. (дата обращения 20.09.2012).
2. Kenny P. Bayesian Analysis of Speaker Diarization with Eigenvoice Priors // Technical report, Centre de recherche informatique de Montreal (CRIM). – Montreal, Canada. – May 2008. – 17 p.
3. Кудашев О.Ю., Пеховский Т.С. Проблема инициализации систем сегментации дикторов на основе вариационного байесовского анализа // Научно-технический вестник информационных технологий, механики и оптики. – 2012. – № 3 (79). – С. 83–87.
4. Reynolds D., Kenny P., Castaldo F. A Study of New Approaches to Speaker Diarization // Proc. Interspeech – 2009. – P. 1047–1050.
5. Jin Q., Laskowski K., Schultz T., Alex Waibel A. Speaker segmentation and Clustering in Meetings // Proc. ICASSP-2004 Meeting Recognition Workshop. – Montreal, Canada. – May 2004. – P. 112–117.
6. Радио Свобода [Электронный ресурс]. – URL: http://www.svobodanews.ru/, свободный. Яз. рус. (дата обращения 20.09.2012).

Кудашев Олег Юрьевич – ООО «ЦРТ-инновации», программист, Санкт-Петербургский национальный исследовательский университет информационных технологий, механики и оптики, аспирант, kudashev@speechpro.com