УДК 004.93+57.087.1

ПРОБЛЕМЫ ИНИЦИАЛИЗАЦИИ СИСТЕМ СЕГМЕНТАЦИИ ДИКТОРОВ НА ОСНОВЕ ВАРИАЦИОННОГО БАЙЕСОВСКОГО АНАЛИЗА

О.Ю. Кудашев, Т.С. Пеховский

Приведено описание модели, используемой для решения задачи сегментации дикторов. На основе сделанных предположений приведены итерационные формулы аппроксимации функции апостериорного распределения параметров модели диктора и предложен оригинальный способ инициализации значений параметров модели. Приведена схема системы сегментации дикторов, реализованной на основе разработанного подхода. Применение разработанной системы дало относительную редукцию ошибки до 26% как на англоязычных, так и русскоязычных речевых базах.
Ключевые слова: байесовский анализ, вариационный метод, сегментация дикторов.

Введение

Задача сегментации дикторов состоит в выделении речевых сегментов фонограммы и кластеризации (объединении) выделенных сегментов по принадлежности к одному диктору. Сегментация дикторов является неотъемлемой частью задач, связанных с обработкой речи. К таким задачам можно отнести, например, автоматическую идентификацию голоса говорящего (диктора), индексацию аудиоданных.

В последнее время методы байесовского факторного анализа показали высокую эффективность как в задачах голосовой идентификации [1, 2], так и в задачах сегментации дикторов [3, 4]. Работа [5] является ярким примером алгоритма сегментации дикторов, основанного на вариационном байесовском анализе. Однако, как и в любом итерационном алгоритме, возникает вопрос о начальной инициализации значений.

Целью данной работы является разработка и применение алгоритма инициализации начальных значений параметров модели, основанной на вариационном байесовском анализе. В отличие от работы [6], исследуется система сегментации дикторов для широкого спектра приложений, в частности, на различных русскоязычных и англоязычных речевых базах.

Применение вариационного байесовского анализа к задаче сегментации дикторов

Пусть X – данные; θ – совместный набор параметров модели и скрытых переменных. Задачей байесовского анализа является поиск максимально точного приближения $Q(\theta)$ для функции апостериорного распределения параметров модели $P(\theta|X)$:

$$P(\theta|X) = \frac{P(X|\theta)P(\theta)}{P(X)},$$

где $P(X) = \int P(X|\theta)P(\theta)d\theta$.

Доказано, что решением данной задачи является функция $Q(\theta)$, обеспечивающая максимум нижней границы $L(Q)$:

$$L(Q) = \int Q(\theta)\ln\frac{P(X,\theta)}{Q(\theta)}d\theta.$$

Наиболее распространенными численными методами решения задачи байесовского анализа являются вариационные байесовские методы [7].

В основе как методов сегментации дикторов, так и методов текстонезависимой голосовой идентификации лежит статистическое моделирование распределения акустических признаков. Наиболее эффективным типом генеративной модели диктора признана смесь гауссовых распределений (Gaussian Mixture Models, GMM), аппроксимирующая распределение акустических признаков. При этом модель каждого диктора получается из универсальной фоновой модели (Universal Background Model, UBM) путем адаптации только средних значений гауссоид без изменений матриц ковариаций. Объединение полученных таким образом векторов средних значений называют супервектором средних диктора.

Введем следующие предположения:

- речевые сегменты на фонограмме выделены, при этом на каждом сегменте присутствует ровно один диктор;
- количество дикторов известно;
- супервектор средних диктора имеет следующее априорное распределение:

$$\mathbf{S} = \boldsymbol{\mu} + \mathbf{V}\mathbf{y},$$

где $\boldsymbol{\mu}$ – супервектор средних UBM; $\mathbf{V}$ – матрица «собственных голосов»; $\mathbf{y}$ – случайный вектор с нормальным стандартным распределением.

Последнее предположение является основой факторного анализа и дает не только мощный инструмент для методов сегментации дикторов, но также значительно уменьшает вычислительную сложность алгоритмов.

Введем следующие обозначения:

- M – число речевых сегментов на фонограмме;
- S – число дикторов;
- $X = (\mathbf{x}_1, ..., \mathbf{x}_M)$ – данные, полученные на каждом из сегментов фонограммы;
- $I = (\mathbf{i}_1, ..., \mathbf{i}_M)$ – набор индикаторов для каждого сегмента фонограммы (i_{ms}=1, если на сегменте m говорит диктор s и равен нулю в обратном случае);
- $\pi = (\pi_1, ..., \pi_M)$ – априорные вероятности присутствия диктора s на сегменте;
- $Y = (\mathbf{y}_1, ..., \mathbf{y}_S)$ – вектора в пространстве собственных голосов, отвечающие моделям соответствующих дикторов;
- $P(\mathbf{i}_m) = \prod_{s=1}^{S} \pi_s^{i_{ms}}$ – априорное распределение $\mathbf{i}_m$;
- $P(Y)$ имеет нормальное стандартное распределение.

В соответствии с вариационным байесовским анализом, сделаем следующее предположение о факторизации:

$$Q(Y, I) = Q(Y)Q(I),$$

$$Q(I) = \prod_{m=1}^{M}\prod_{s=1}^{S} q_{ms}^{i_{ms}},$$

$$Q(Y) = \prod_{s=1}^{S} N\left(\mathbf{y}_s \middle| \mathbf{a}_s, \mathbf{\Lambda}_s^{-1}\right).$$

Тогда формулы для вариационного приближения искомых функций имеют следующий общий вид:

$$\ln Q(Y) = E_I\left[\ln P(X, Y, I\pi)\right] + \text{const}, \quad (1)$$

$$\ln Q(I) = E_Y\left[\ln P(X, Y, I\pi)\right] + \text{const}. \quad (2)$$

Подробный вывод формул представлен в работе [5].

Таким образом, применяя последовательно формулы (1), (2), можно получить приближенные значения q_{ms}, являющиеся апостериорной вероятностью присутствия диктора s на сегменте m.

Система сегментации дикторов

Алгоритм вариационного байесовского анализа, вообще говоря, гарантирует сходимость только в окрестности локального максимума функции $L(Q)$. Следовательно, для эффективного применения алгоритма, описанного в предыдущем разделе, необходимо произвести удачную инициализацию начальных значений q_{ms}.

Авторами была рассмотрена задача сегментации дикторов при S=2 (диалог). Инициализация начальных значений q_{m1}, q_{m2} осуществляется путем применения алгоритма K-средних для векторов $\mathbf{y}_m$ каждого из сегментов фонограммы с дальнейшей кластеризацией. Подробное описание такой кластеризации приведено в работе [6].

Общая схема взаимодействия блоков разработанной системы сегментации дикторов представлена на рисунке. Такая система состоит из шести основных блоков.

1. блок выделения речевых сегментов (Voice Activity Detector, **VAD**);
2. блок выделения речевых акустических признаков (Feature Extractor, **FE**);
3. блок построения векторов $\{\mathbf{y}_m\}_{m=1}^{M}$ для каждого сегмента фонограммы (***Y*-mapping**);
4. блок кластеризации множества векторов $\{\mathbf{y}_m\}_{m=1}^{M}$ на множества C_1, C_2 при помощи алгоритма K-средних (***K*-means**);
5. блок перегруппировки множеств C_1, C_2 в соответствии с формулой (**EV)**

$$C_1^* = \left\{m : (<\mathbf{y}>_1 - <\mathbf{y}>_2)\cdot(\mathbf{y}_m - <\mathbf{y}>_1) > 0\right\},$$

$$C_2^* = \left\{m : (<\mathbf{y}>_2 - <\mathbf{y}>_1)\cdot(\mathbf{y}_m - <\mathbf{y}>_2) > 0\right\},$$

где $<\mathbf{y}>_1 = \frac{1}{|C_1|}\sum_{m\in C_1}\mathbf{y}_m$; $<\mathbf{y}>_2 = \frac{1}{|C_2|}\sum_{m\in C_2}\mathbf{y}_m$;

6. блок перегруппировки множеств C_1, C_2 на основе вариационного байесовского анализа (**VBA**).

Инициализация начальных значений q_{m1}, q_{m2} осуществляется следующим образом:

$$q_{ms} = \begin{cases} 0{,}999\, m \in C_s \\ rand()\, m \notin C_s \end{cases}.$$

Окончательная группировка множеств C_1, C_2 осуществляется после применения вариационного байесовского анализа в соответствии с формулой $C_1^* = \{m : q_{m1} > q_{m2}\}, C_2^* = \{m : q_{m2} \geq q_{m1}\}$.

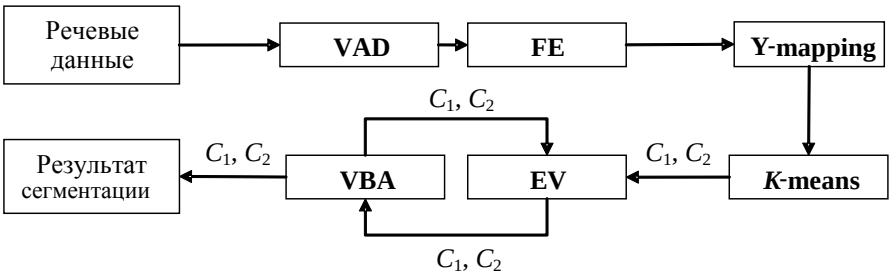

Рисунок. Схема взаимодействия блоков системы сегментации дикторов

Таким образом, происходит взаимодействие между дискриминативным (**EV**) и генеративным (**VBA**) блоками [6].

Результаты экспериментов

Основной метрикой эффективности системы сегментации дикторов является показатель вероятности ошибки сегментации (Diarization Error Rate, DER):

$$DER = \frac{\text{длина речевых сегментов, неверно отнесенных к диктору}}{\text{длина всех речевых сегментов}} 100\% .$$

Помимо описанной системы, в экспериментах также была применена схема, отличающаяся отсутствием блоков **Y-mapping**, ***K*-means** и **EV**. При этом инициализация блока **VBA** происходило 6 раз путем рандомизации значений q_{m1}, q_{m2}. Из шести полученных множеств C_1, C_2 выбирались те, которые обеспечивали наибольшее значение нижней границы $L(Q)$.

В качестве речевых акустических признаков использовались первые 13 MFCC (Mel-Frequency Cepstral Coefficients) коэффициентов, без нормализации и вычитания среднего значения. Для обучения UBM и матрицы собственных голосов **V** использовались следующие речевые базы: NIST 2002, NIST 2003, NIST 2004, NIST 2005, NIST 2006, NIST 2008, RuSTeN [8]. Суммарное количество дикторов указанных баз обучения составило 3620. Количество гауссовых распределений UBM бралось равным 512. Размерность матрицы собственных голосов для блоков ***K*-means** и **EV** составила 10, а размерность матрицы собственных голосов для блока **VBA** – 50.

В таблице представлены сравнительные результаты экспериментальных исследований реализованной системы сегментации дикторов (столбец DER), системы, основанной на рандомизации значений q_{m1}, q_{m2} (столбец DER random), а также процент относительной редукции. В тестировании участвовали как англоязычные базы данных (NIST 2002, NIST 2008), так и русскоязычные базы компании ООО «Центр речевых технологий» (НАРКОКОНТРОЛЬ, МВД, СУБТИТРЫ). Базы НАРКОКОНТРОЛЬ и МВД представляют собой записи телефонных разговоров длительностью 1–3 мин. База СУБТИТРЫ состоит из коротких записей (20–50 с) новостных интервью, включающих разговор двух дикторов.

Наименование базы данных	Число файлов	Язык	Канал	Средняя продолжительность записей	DER, %	DER random, %	Редукция, %
NIST 2002	88	англ.	телефон	1 мин	5,15	7,03	26
NIST 2008	1172	англ.	телефон	5 мин	5,55	5,86	5,3
НАРКОКОНТРОЛЬ	17	русский	телефон	1 мин 45 с	4,8	4,97	3,4
МВД	83	русский	телефон	2 мин	5,77	5,95	3
СУБТИТРЫ	103	русский	радио	30 с	4,53	4,75	4,6

Таблица. DER систем сегментации дикторов

Заключение

Как видно из таблицы, применение предварительной инициализации блоком **EV** обеспечило относительную редукцию DER на всех тестируемых базах. При этом необходимо отметить, что результаты справедливы как для англоязычных, так и для русскоязычных корпусов. На англоязычной базе с короткими произнесениями (NIST 2002) относительная редукция DER составила 26%. Разброс величины редукции обусловлен, главным образом, разнообразными условиями тестирования. Помимо усиления системы сегментации дикторов, использование дискриминативного блока **EV** имеет еще одно важное прак-

тическое значение. Разработанная авторами схема позволяет свести к минимуму число обращений к блоку **VBA** (в среднем 2–3 раза). Поскольку этот блок имеет наибольшую вычислительную сложность, происходит значительное уменьшение времени работы всей системы.

Разработанная система успешно внедрена и используется в системах автоматической голосовой идентификации, разработанных на кафедре «Речевые информационные системы», являющейся базовой кафедрой компании ООО «Центр речевых технологий». Также указанная система имела успешное применение в составе системы автоматической индексации записей новостных передач.

Литература

1. Kenny P., Ouellet P., Dehak N., Gupta V., Dumouchel P. A study of inter-speaker variability in speaker verification // IEEE Trans. Audio, Speech and Lang. Process. – July 2008. – V. 16. – № 5. – P. 980–988.
2. Castaldo F., Colibro D., Dalmasso E., Laface P., Vair C. Compensation of Nuisance Factors for Speaker and Language Recognition // IEEE Trans. Audio, Speech, Lang. Process. – September 2007. – V. 15. – № 7. – P. 1969–1978.
3. Tranter S., Reynolds D. An overview of automatic speaker diarization systems // IEEE Trans. Audio, Speech, Lang. Process. – September 2006. – V. 14. – № 5. – P. 1557–1565.
4. Reynolds D., Kenny P., Castaldo F. A Study of New Approaches to Speaker Diarization // Proc. Interspeech – 2009. – P. 1047–1050.
5. 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.
6. Пеховский Т.С., Шулипа А.К. Гибрид генеративных и дискриминативных моделей для задачи диаризации в коротком телефонном диалоге // Proc. of the XIV International Conference «Speech and Computer» SpeCom'11. – Kazan, Russia, 2011. – P. 389–394.
7. Bishop M. Pattern Recognition and Machine Learning. – New York: Springer, 2006. – 738 p.
8. Linguistic Data Consortium [Электронный ресурс]. – URL: http://www.ldc.upenn.edu/, свободный. Яз. англ. (дата обращения 10.03.2012).

Кудашев Олег Юрьевич – Санкт-Петербургский национальный исследовательский университет информационных технологий, механики и оптики, аспирант, kudashev@speechpro.com

Пеховский Тимур Сахиевич – ООО «ЦРТ-инновации», кандидат физ.-мат. наук, ведущий научный сотрудник, tim@speechpro.com