А. Л. Ронжин, В. Ю. Будков

АНАЛИЗ СОВРЕМЕННЫХ МЕТОДОВ И СИСТЕМ ДИАРИЗАЦИИ ДИКТОРОВ

Рассматривается проблема диаризации (протоколирования) речи нескольких дикторов, записанной одно- или многоканальными аудиосистемами. Проанализированы современные подходы к решению проблемы и приведены методики оценивания эффективности работы систем диаризации.

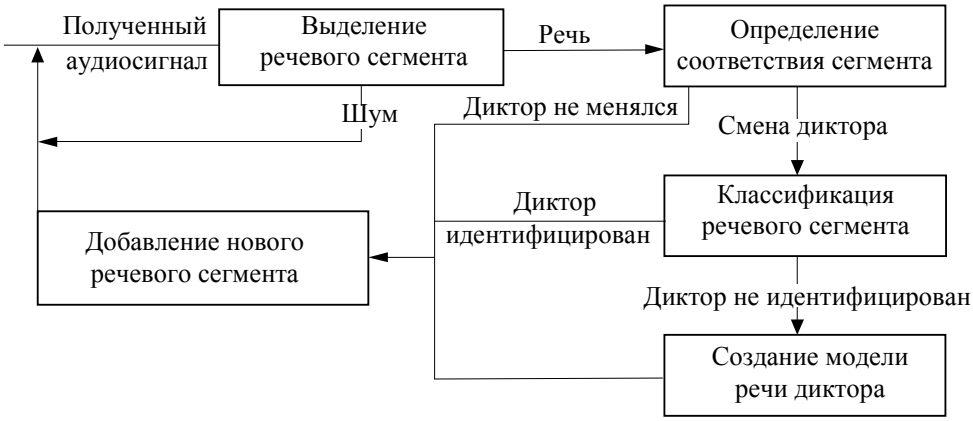
Ключевые слова: *цифровая обработка аудиосигнала, протоколирование речи дикторов, уровень ошибок диаризации.*

Введение. Задача протоколирования речи дикторов (Speaker Diarization — SD), также известная в зарубежной литературе под названием “Who Spoke When” (кто и когда говорил), заключается в сегментации входного звукового сигнала по типу аудиоинформации и его источнику [1—3]. Аудиосигнал может содержать речь диктора, одновременную речь нескольких дикторов, музыку, фоновые шумы. Наиболее перспективными областями применения систем диаризации дикторов являются:

— системы аннотирования, добавляющие к речевым аудиофайлам различные метаданные, такие как временная разметка границ фраз, информация о говорящем: это позволяет ускорить „ручной“ поиск данных и упростить их автоматизированную обработку;

— системы автоматического распознавания речи, использующие диаризацию дикторов для адаптации моделей фонем к речи пользователя, что повышает точность распознавания речи.

Структура типовой системы диаризации. Процесс протоколирования речи дикторов состоит из двух основных этапов: сегментации реплик каждого диктора в аудиосигнале и последующей группировки всех сегментов по принадлежности к каждому из дикторов [2]. Структура типовой системы диаризации дикторов приведена на рисунке.



Вначале определяются фрагменты, содержащие паузы или шумы, и выделяются границы речевого сегмента. Полученный речевой сегмент используется для определения (проверки) его

принадлежности (или отсутствия таковой) соответствующей модели речи текущего диктора. Если соответствие установлено, то выделенный речевой сегмент добавляется к аудиофайлу данного диктора. В случае замены диктора производится классификация сегмента по принадлежности к существующим моделям речи дикторов или создается модель речи нового диктора и сегмент помечается как принадлежащий последнему. Для повышения точности работы систем диаризации при обработке речевых сегментов могут приниматься во внимание дополнительные параметры — качество сигнала и пол диктора [2].

Рассмотрим основные методы обработки аудиосигнала, применяемые на каждом из уровней, указанных в структуре типовой системы диаризации.

Методы обработки звукового сигнала. Предварительное выделение из аудиосигнала фрагментов, содержащих тишину или речь, позволяет значительно сократить уровень ошибок системы диаризации и повысить скорость обработки данных. Методы определения речевой активности (Voice Activity Detection — VAD), основанные на оценке уровня энергии сигнала или его спектра и хорошо зарекомендовавшие себя при обработке речи, записанной с помощью одного микрофона, однако, не решают проблем, возникающих при обработке многоканальных аудиозаписей мероприятий с участием нескольких дикторов [4]. Для решения этой проблемы используются методы, основанные на нормализации энергии многоканального сигнала [5] и оценке степени корреляции между каналами [6], а также скрытые марковские модели, содержащие не два состояния (речь/тишина), как обычно в VAD-методах, а $2K$ состояний, где K — количество дикторов [3]. Особенностью этих моделей является необходимый предварительный этап их обучения. Применение корреляционных методов возможно только при обеспечении синхронности многоканальной записи аудиопотоков.

Большинство современных систем протоколирования при построении моделей речи дикторов используют гауссовы смеси (Gaussian Mixture Models — GMMs). Данный подход позволяет обеспечить достаточно высокую точность работы систем, но требует предварительного обучения моделей, что ограничивает возможность применения этого метода в режиме реального времени. В работе [7] предложена оригинальная методика использования высокопроизводительных многоядерных процессоров параллельного вычисления, позволившая реализовать обучение моделей GMM.

При протоколировании речи дикторов наиболее часто применяется метод параметрического представления звукового сигнала с использованием признаков MFCC (Mel Frequency Cepstral Coefficients), успешно зарекомендовавший себя также в системах распознавания речи [2]. В работе [8] предложены новые спектральные признаки, которые сравниваются с традиционными MFCC-признаками. В ходе экспериментов установлено, что применение новых признаков приводит к снижению ошибки диаризации на 15,4 %. Для протоколирования речи дикторов также дополнительно использовалась система локализации источника звука, что обеспечило снижение ошибки диаризации на 5,2 % [8].

Протоколирование мероприятий обычно основывается только на акустической информации [9]. Однако роль участников мероприятия и последовательность их выступлений связаны и статистически предсказуемы. В работе [10] описывается использование моделей N-грамм ролей участников для определения структуры (шаблона) дискуссии и применение этой структуры в работе системы протоколирования. Предложенный метод позволяет уменьшить ошибки системы протоколирования на 19 %, когда роль участника заранее известна, и на 17 %, когда она определяется системой в ходе обработки аудиозаписи [10].

Методики оценивания эффективности работы систем диаризации. Одна из основных методик оценивания систем диаризации [1] основана на вычислении уровня ошибок диаризации (Diarization Error Rate — DER). Данная метрика показывает, какова длительность фрагмента аудиосигнала, просегментированного некорректно, т.е. с ошибками при идентифи-

фикации диктора или при определении шума/речи. Показатель DER вычисляется по следующей формуле:

$$\mathrm{Error}_{\mathrm{SpkrSeg}} = \frac{\sum_{\mathrm{all\,segs}}\{\mathrm{dur(seg)}(\max(N_{\mathrm{Ref}}(\mathrm{seg}),\, N_{\mathrm{Out}}(\mathrm{seg}) - N_{\mathrm{Correct}}(\mathrm{seg}))\}}{\sum_{\mathrm{all\,segs}}\{\mathrm{dur(seg)}N_{\mathrm{Ref}}(\mathrm{seg})\}},$$

где $\mathrm{dur(seg)}$ — длительность некоторого речевого сегмента; $N_{\mathrm{Ref}}(\mathrm{seg})$ — количество дикторов, речь которых присутствовала в этом сегменте; $N_{\mathrm{Out}}(\mathrm{seg})$ — количество дикторов, определенное системой диаризации; $N_{\mathrm{Correct}}(\mathrm{seg})$ — количество дикторов, речь которых была записана и которые корректно распознаны системой диаризации.

В приведенной формуле можно выделить следующие составляющие: $\mathrm{Error}_{\mathrm{Spkr}}$ — длительность речевого сигнала, выделенного с ошибками при идентификации диктора:

$$\mathrm{Error}_{\mathrm{Spkr}} = \frac{\sum_{\mathrm{all\,segs}}\{\mathrm{dur(seg)}(\min(N_{\mathrm{Ref}}(\mathrm{seg}),\, N_{\mathrm{Out}}(\mathrm{seg}) - N_{\mathrm{Correct}}(\mathrm{seg}))\}}{\sum_{\mathrm{all\,segs}}\{\mathrm{dur(seg)}N_{\mathrm{Ref}}(\mathrm{seg})\}};$$

также точность сегментации аудиопотока с учетом индивидуальных особенностей речи дикторов можно оценить по числу ложных (False Alarm — FA) и пропущенных (Miss Rate — MS) сегментов речи; длительность ложных речевых сегментов (FA-rate) оценивается как

$$\mathrm{Error}_{\mathrm{FA}} = \frac{\sum_{\mathrm{all\,segs}}\{\mathrm{dur(seg)}(N_{\mathrm{Out}}(\mathrm{seg}) - N_{\mathrm{Ref}}(\mathrm{seg}))\}}{\sum_{\mathrm{all\,segs}}\{\mathrm{dur(seg)}N_{\mathrm{Ref}}(\mathrm{seg})\}},$$

а длительность пропущенных речевых сегментов (MS-rate) — как

$$\mathrm{Error}_{\mathrm{MS}} = \frac{\sum_{\mathrm{all\,segs}}\{\mathrm{dur(seg)}(N_{\mathrm{Ref}}(\mathrm{seg}) - N_{\mathrm{Out}}(\mathrm{seg}))\}}{\sum_{\mathrm{all\,segs}}\{\mathrm{dur(seg)}N_{\mathrm{Ref}}(\mathrm{seg})\}}.$$

Следует отметить, что, кроме перечисленных выше, существуют и другие методики оценивания систем диаризации, эффективность которых исследуется в настоящее время.

Заключение. Рассмотренные методы и системы диаризации дикторов различаются по скорости работы, необходимости предварительного обучения систем, точности сегментации аудиосигнала. Большинство разрабатываемых систем используют признаки MFCC для параметрического представления звукового сигнала; применение дополнительной информации (не только акустической) и распределенных микрофонов позволяет учесть индивидуальные особенности речи дикторов и улучшить точность работы систем диаризации.

Статья подготовлена по результатам исследований, проводимых при поддержке Минобрнауки РФ (федеральная целевая программа „Исследования и разработки“, госконтракт № 07.514.11.4139) и Российского фонда фундаментальных исследований (проект № 12-08-01265-а).

СПИСОК ЛИТЕРАТУРЫ

1. The US NIST 2009, Rich Transcription Evaluation [Электронный ресурс]: <http://www.itl.nist.gov/iad/894.01/tests/rt/2009>.

2. *Tranter S., Reynolds D.* An overview of automatic speaker diarization systems // IEEE Trans. ASLP. 2006. Vol. 14, N 5. P. 1557—1565.

3. *Будков В. Ю., Прищепа М. В., Ронжин А. Л., Марков К.* Многоканальная система анализа речевой активности участников совещания // Третий междисциплинарный семинар „Анализ разговорной русской речи“: Тр. СПИИРАН. 2009. С. 57—62.

4. *Pfau T., Ellis D., Stolcke D.* Multispeaker speech activity detection for the ICSI meeting recorder // Proc. IEEE ASRU Workshop. Madonna di Campiglio, Italy, 2001. P. 107—110.

5. *Dines J., Vepa J., Hain T.* The segmentation of multi-channel meeting recordings for automatic speech recognition // Proc. Interspeech-2006, Pittsburgh, PA. 2006. P. 1213—1216.

6. *Flego F., Zieger C., Omologo M.* Adaptive weighting of microphone arrays for distant-talking F0 and voiced/unvoiced estimation // Proc. Interspeech-2007, Antwerpen, Belgium. 2007. P. 2961—2964.

7. *Qiao Li, Qing Fan, Yunpeng Xiao, Weiping Ye.* A comparable study on PNCC in speaker diarization for meetings // Proc. of the 1st ACIS Intern. Symp. on Cryptography and Network Security, Data Mining and Knowledge Discovery, E-Commerce & Its Applications and Embedded Systems (CDEE 2010), Yanshan Univ., China. 2010. P. 157—160.

8. *Zhou Yu, Suo Hongbin, Wang Junjie, Yan Yonghong.* An improved speaker diarization system for multiple distance microphone meetings // Proc. of the 5th Intern. Conf. on Intelligent Computation Technology and Automation (ICICTA 2012), Zhangjiajie, Hunan. 2012. P. 80—83.

9. *Ронжин А. Л., Будков В. Ю.* Технологии поддержки гибридных е-совещаний на основе методов аудио-визуальной обработки // Вестн. компьютерных и информационных технологий. 2011. № 4. С. 31—35.

10. *Valente F., Vijayasenan D., Motlicek P.* Speaker diarization of meetings based on speaker role n-gram models // Proc. IEEE ASRU Workshop. Madonna di Campiglio, Italy, 2011. P. 4416—4419.

Сведения об авторах

Андрей Леонидович Ронжин — д-р техн. наук, доцент; СПИИРАН, лаборатория речевых и многомодальных интерфейсов; E-mail: ronzhin@iias.spb.su

Виктор Юрьевич Будков — аспирант; СПИИРАН, лаборатория речевых и многомодальных интерфейсов; E-mail: budkov@iias.spb.su

Рекомендована СПИИРАН

Поступила в редакцию
10.06.12 г.