Система протоколирования дикторов на базе алгоритма определения речевой активности в многоканальном аудиопотоке

Ронжин А.Л.
Будков В.Ю.

Учреждение Российской академии наук Санкт-Петербургский институт информатики и автоматизации РАН.
Россия, 199178 Санкт-Петербург, 14 линия, д. 39.
Тел.: (812) 328-7081; Факс: (812) 328-7081.
E-mail: {ronzhin, budkov}@iias.spb.su

Рассматривается система многоканальной записи и последующего анализа речи участников мероприятий в интеллектуальном зале. Предложен комплекс алгоритмов для выделения границ фраз в многоканальном аудиопотоке, записанном встроенными микрофонами веб-камер, расположенных на конференц-столе перед каждым участником. Выбранный в ходе экспериментов алгоритм позволяет достичь приемлемого качества выделения границ фраз и автоматически выбирать номер камеры текущего активного диктора.

Интернет-приложения для телеконференций и дистанционного обучения, так называемые системы E-meeting и E-lecture, становятся всё более популярными в коммерческих, исследовательских, образовательных и других организациях. Такие системы позволяют сэкономить на транспортных расходах, выбрать индивидуальный способ обучения, а также предоставляют удобные средства поиска и доступа к информации. Тем не менее, большая часть работы по протоколированию, ведению хода мероприятия, подключению отдельных удалённых участников выполняется вручную оператором или секретарём. Задача протоколирования дикторов (speaker diarization (SD)), также известная в зарубежной литературе под названием «Who Spoke When», состоит в сегментации реплик каждого диктора в аудиосигнале и последующей группировке всех сегментов каждого диктора [1]. В процессе протоколирования SD системы выполняют ряд последовательных операций [2]. Вначале определяются границы речи и участки, содержащие паузы или шумы, затем проверяется, изменился ли текущий диктор, определяется пол диктора и наконец производится классификация сегмента речи среди существующих дикторов или создается модель нового диктора.

Предварительная сегментация сигнала на участки, содержащие тишину или речь, позволяет значительно сократить уровень ошибок распознавания речи, повысить скорость обработки. К сожалению, методы определения речевой активности (voice activity detection (VAD), основанные на оценке уров-

ня энергии сигнала или его спектра, хорошо зарекомендовавшие себя при обработке речи, записанной с помощью одного микрофона, не решают проблем, возникающих при обработке многоканальных аудиозаписей мероприятий с несколькими дикторами [3]. Для решения этой проблемы используются методы, основанные на нормализации энергии многоканального сигнала [4], оценке степени корреляции между каналами [5], а также скрытые макровские модели, содержащие не 2 состояния (речь/тишина), как обычно в VAD методах, а 2^K состояний, где K — число дикторов [6]. Их особенностью является необходимый предварительный этап обучения моделей, поэтому на данной стадии исследования для определения речи в многоканальной системе были использованы более простые подходы, выполняющие классификацию без настройки моделей. Применение корреляционных методов возможно только при обеспечении синхронности многоканальной записи аудиопотоков. В случае же распределённых мероприятий и использования независимых устройств записи и обработки аудиосигналов более эффективно применение методов на основе нормализации энергии сигналов в аудиоканалах, расчёта относительной энергии сегмента и его спектра, учёта фонетических закономерностей речи.

В данной работе описаны результаты исследования и разработки системы многоканальной записи и последующего анализа речи участников мероприятий в интеллектуальном зале. Для записи поведения участников и последующего выделения в аудио- и видеосигналах сегментов, содержащих речь, жесты или другую активность, связанную с ходом мероприятия, были использованы веб-камеры Logitech Sphere AF со встроенным микрофоном. Более полное описание оборудования и программных средств, использованных при разработке интеллектуального зала, можно найти в [7].

Обычно участники сидят достаточно близко друг к другу за столом совещаний, поэтому соседние микрофоны могут захватывать речь одного и того же диктора с примерно одинаковой амплитудой сигнала. В итоге определение границ речи по энергии сигнала или его спектра (в каждом канале независимо) часто приводит к ошибочным результатам. Для повышения точности анализа применяют различные способы нормализации [2]. В работе [4] рассчитывается относительная энергия сегмента сигнала E_n^{norm} в каждом канале:

$E_n^{norm}(i) = \frac{E_n(i)}{\sum_{k=1}^{M} E_k(i)}$, где $E_n(i)$ - энергия в канале n для сегмента i, M - число каналов в системе. Таким образом, нормализованная энергия сегмента для каждого канала будет рассчитана относительно всех каналов в системе, и её значение будет изменяться в диапазоне от нуля до единицы. Для компенсации различий в усилении сигнала по разным каналам в работе [3] было предложено дополнительно учитывать минимальную энергию сегмента в каждом канале: $E_n^{norm}(i) = \log_0 (E_n(i) - E_n^{\min} - \frac{1}{M} \cdot \sum_{k=1}^{M} E_k(i))$, где $E_n^{\min}$ — минимальная энергия сегмента, вычисленная для каждого канала в условиях тишины, вычитание которой позволяет учесть различные уровни усиления и внутренние шумы микрофонов. Затем после вычитания средней энергии по каналам производится логарифмирование, чтобы сократить разрядность полученного значения энергии. Для этой же цели используется расчёт коэффициентов усиления по каждому каналу: $E_n^{norm}(i) = \log_0 (E_n(i) \cdot \kappa_{Amp}^n - \frac{1}{M} \cdot \sum_{k=1}^{M} \kappa_{Amp}^k \cdot E_k(i))$, где κ_{Amp}^n — коэффициент усиления n канала, который позволяет учесть различные уровни записи микрофонов. Нормализованная энергия $E_n^{norm}(i)$ показывает относительное усиление сигнала в каждом канале и позволяет определить наличие речи в текущем сегменте. Последний алгоритм (Relative Energy Estimation (*REE*)) был экспериментально проверен в ходе исследований.

В предложенном алгоритме *RESW* (Relative Energy estimation in Sliding Window) текущий активный диктор (и соответствующий номер веб-камеры) $\hat{\omega}_t$ в момент времени t определялся путём расчёта относительной энергии канала в скользящем окне, за счёт чего подавлялись случайные всплески энергий в отдельных каналах:

$$\hat{\omega}_t = \arg\max_n [\log_0 \{\frac{1}{N} \cdot (\sum_{i=0}^{N-1} \kappa_{Amp}^n \cdot E_n(t+i) - \frac{1}{M} \cdot \sum_{j=1}^{M} \sum_{i=0}^{N-1} \kappa_{Amp}^j \cdot E_j(t+i)\}],$$

где N — размер скользящего окна (число сегментов), M — число аудиоканалов каналов, $E_n(t+i)$ — кратковременная энергия сегмента речи:

$$E_n(t+i) = \sum_{j=0}^{L-1} x^2_{L(t+i)+j}.$$

Для повышения робастности предложенного алгоритма определения речевой активности для выбранного канала $\hat{\omega}_t$ был использован дополнительный анализ, в ходе которого оценивался показатель W, равный числу сегментов в скользящем окне длиной N, значение энергии $E_{\hat{\omega}_t}$ которых превышало заданный порог E_{sil}: $W = \sum_{i=0}^{N-1} f(t,i)$ где $f(t,i) = \begin{cases} 0, E_{\hat{\omega}_t}(t+i) \le E_{sil} \\ 1, \quad иначе \end{cases}$.

Применение пороговой функции $f(t,i)$ позволяет предварительно классифицировать аудиосегмент как тишина ($f(t,i)=0$) или речь ($f(t,i)=1$). Были предложены два варианта оценивания показателя W. Решение о наличии речи в текущем скользящем окне в канале $\hat{\omega}_t$ принималось только в том случае, если показатель W: а) был больше нуля (алгоритм $RESW_1$), либо б) превышал некоторое значение W_{sil} (алгоритм $RESW_{sil}$). Для подавления единичных ложных речевых сегментов учитывалась максимально допустимая пауза $d_{\max}$ между речевыми сегментами. Если число сегментов тишины между текущим речевым сегментом и ближайшим слева или справа сегментом речи в скользящем окне превышало значение $d_{\max}$, то текущий сегмент речи классифицировался как тишина: $W^{sil} = \sum_{i=0}^{N-1} [f(t,i) \wedge \phi(t,i)]$,

где $\phi(t,i) = \begin{cases} 1, \left[\sum_{l=i-d_{\max}}^{i+d_{\max}} f(t+l,i)\right] - 1 > 0 \\ 0, \quad иначе \end{cases}$. В данном алгоритме ($RESW_{sil+dist}$) выполнение условия $W^{sil} > W_{sil}$ определяло наличие речи в скользящем окне.

В таблице представлено краткое описание четырёх алгоритмов, которые были использованы при тестировании и выборе оптимального способа определения границ речи в многоканальном аудиопотоке.

Таблица. Алгоритмы определения речевой активности в многоканальном аудиопотоке

Обозначение алгоритма	Описание
REE	Сравнение относительной энергии сегментов в аудиоканалах с порогом E_{sil}.
$RESW_1$	Проверка наличия хотя бы одного сегмента в скользящем окне, значение энергии которого превышает порог E_{sil}.
E_{sil}	Сравнение числа сегментов в скользящем окне, значения энергий которых превышает порог E_{sil}, с максимально допустимым для тишины порогом W_{sil}.

$RESW_{sil+dist}$	Сравнение числа сегментов в скользящем окне, значения энергий которых превышает порог E_{sil}, с порогом W_{sil} и учет максимально допустимой паузы d_{max} между речевыми сегментами.

Точность сегментации аудиопотока по дикторам оценивалась по числу ложных (false alarm (*FA*)) и пропущенных (miss rate (*MS*)) сегментов речи. При анализе работы многоканальной системы оценки суммируются по всем каналам *M* [8]:

$MS = \sum_{k=1}^{M} T_k^{(MS)} / (\sum_{k=1}^{M} T_k^{(S)} + \sum_{k=1}^{M} T_k^{(MS)})$, $FA = \sum_{k=1}^{M} T_k^{(FA)} / (\sum_{k=1}^{M} T_k^{(S)} + \sum_{k=1}^{M} T_k^{(FA)})$, где $T_k^{(S)}$ — число сегментов речи в канале k, верно определенных системой как речь; $T_k^{(MS)}$ — число сегментов речи, пропущенных системой; $T_k^{(FA)}$ — число неречевых сегментов, определенных системой как речь. При настройке параметров алгоритма определения границ речи приходится выбирать некоторый компромисс между числом пропущенных и ложных сегментов [1]. Для этой цели служит общепринятая функция DET (detection error trade-off), которая показывает, как зависит уровень пропущенных сегментов речи MS от уровня ложных речевых сегментов FA. С помощью данной зависимости вычисляют коэффициент равных уровней MS и FA (EER — Equal Error Rate) — точка на кривой DET, где значения MS и FA имеют наиболее близкие значения.

Для экспериментальной проверки алгоритмов определения речевой активности была подготовлена тестовая база данных, содержащая пятиканальную аудиозапись с частотой дискретизации отсчетов 16кГц. Длина сегмента речи равнялась 1600 отсчетам. Длина скользящего окна составляла 10 сегментов. Окно сдвигалось с шагом равным одному сегменту. Общая длительность речевого сигнала в базе данных составила 28 минут. В ходе эксперимента пять участников последовательно читали предложения различной длины из одного текста. Распечатанные листы бумаги с текстом лежали на столе перед каждым участником. Таким образом, в данном эксперименте была создана несколько искусственная ситуация: участники не перебивали друг друга, а читали предложения последовательно; между микрофоном и участником не возникали помехи (руки, бумаги, другие предметы), лицо диктора было направлено преимущественно в сторону микрофона на протяжении всей записи.

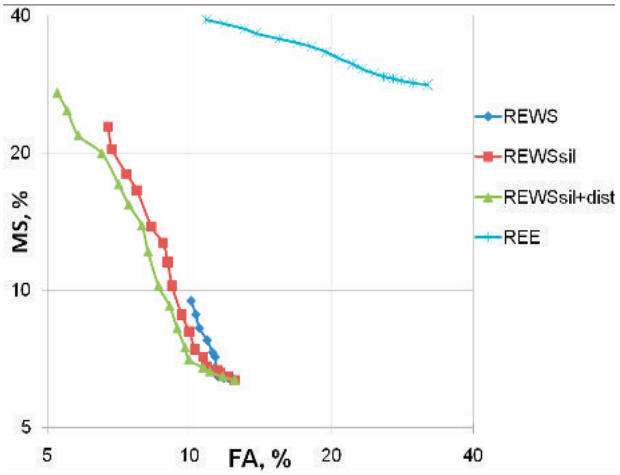

Рис. Уровень ошибок *MS* и *FA* для алгоритмов многоканальной оценки речевой активности

В ходе прослушивания всех записей вручную были выставлены границы фраз в каждом канале с точностью до одного аудиосегмента. Полученная разметка использовалась в качестве эталонной, по которой оценивалось качество автоматической сегментации. На рисунке показано как изменялся уровень ошибок MS и FA при нескольких значениях E_{sil} для алгоритмов REE, $RESW_1$, $RESW_{sil}$ и $RESW_{sil+dist}$. Характер полученных зависимостей согласуется с результатами аналогичных исследований. В данном экс-

перименте с помощью алгоритма $RESW_{sil+dist}$ границы фраз участников были определены наиболее точно ($EER_{RESW_{sil+dist}} = 9{,}16\%$).

Разработка многоканальной системы анализа речевой активности, использующейся при создании мультимедийных отчётов распределённых мероприятий, способствует сокращению трудозатрат при подготовке стенограмм, повышению качества проведения телеконференций и позволяет вести мониторинг и расчёт статистики хода совещания, а также организовать быстрый поиск по мультимедийным архивам. Применение персональных петличных микрофонов в большинстве случаев обеспечивает высокое качество записи, но требует предварительной установки и ограничивает движения диктора. В разработанной системе протоколирования используется набор персональных веб-камер со встроенными микрофонами и алгоритм определения речевой активности в многоканальном аудиопотоке, позволяющих достичь приемлемого качества выделения фраз дикторов и автоматически выбирать камеру участника, активного в текущий момент.

Работа выполнена в рамках ФЦП «Научные и научно-педагогические кадры инновационной России» (ГК №П2360) и грантов РФФИ (№ 08-08-00128-а, 08-07-90002-СТ_а).

ЛИТЕРАТУРА

1. NIST, Rich Transcription 2009 Evaluation, 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, vol.14, no. 5, 2006. P. 1557–1565.

3. *Pfau T., Ellis D., Stolcke D*. Multispeaker Speech Activity Detection for the ICSI Meeting Recorder. In: IEEE ASRU Workshop, 2001. P. 107–110.

4. *Dines J., Vepa J., Hain T*. The segmentation of multi-channel meeting recordings for automatic speech recognition, In: ICSLP-2006. P. 1213–1216.

5. *Flego F., Zieger C., Omologo M*. Adaptive weighting of microphone arrays for distant-talking F0 and voiced/unvoiced estimation. In: Interspeech-2007, 2007. P. 2961–2964.

6. *Laskowski K., Schultz T*. Simultaneous multispeaker segmentation for automatic meeting recognition. In Proc. of EUSIPCO, Poznan, Poland, September 2007. P. 1294–1298.

7. *Будков В.Ю., Прищепа М.В., Ронжин А.Л., Марков К*. Многоканальная система анализа речевой активности участников совещания. Труды третьего междисциплинарного семинара «Анализ разговорной русской речи» (АР^3 – 2009). СПб.: ГУАП, 2009. С. 57–62.

8. *Laskowski K., Jin Q., Schultz T*. Crosscorrelation based multispeaker speech activity detection. In: Interspeech-2004, 2004, Jeju Island, South Korea. P. 973–976.