Сравнение различных способов оценки схожести распределений частоты основного тона в задаче идентификации диктора по его речи

Григорян Р.Л.
Коршунов С.С.
Репалов С.А.
Хрящев М.Ю.

ФГНУ НИИ «Спецвузавтоматика»
344007 Ростов-на-Дону, Газетный пер., д. 51.
Тел. (863) 201-28-17. E-mail: asni@asni.rsu.ru

В настоящей работе рассматриваются результаты исследования по сравнению различных способов оценки удалённости распределений частоты основного тона в задаче идентификации диктора. Как правило, при идентификации дикторов по распределению основного тона решающее правило строится на основе некоторой элементарной оценки схожести. В работе проводится сравнение как элементарных методов оценки схожести гистограмм распределения частоты основного тона, таких как Евклидово расстояние, так и таких методов, как расстояние Кульбака-Лейблера и хи-квадрат. Показывается преимущество методов оценки схожести при использовании расстояний Кульбака-Лейблера и хи-квадрат перед используемыми в настоящее время способами.

Задача автоматической текстонезависимой идентификации дикторов по голосу имеет множество применений. Например, доступ к информации о банковском счёте или идентификация и (или) верификация оператора при голосовом управлении.

Одной из основных характеристик голоса диктора является основной тон — F_0. Эта характеристика считается наиболее изученной, и на данный момент существует множество методик получения значения основного тона на участке речи. В работе [1] описан один из способов, который и был использован в данной работе. Для идентификации диктора используются различные статистические характеристик частоты основного тона. Например, среднее значение частоты основного тона, минимальные и максимальные значение. Одной из наиболее часто используемых характеристик является распределение частоты. При использовании этого метода построение модели диктора состоит в оценке закона распределения. Идентификация состоит в оценке степени близости между двумя распределениями, одно из которых получено на этапе обучения, а второе построено по анализируемой записи голоса [2]. Исследованию различных методов оценки схожести распределений, представленных в виде гистограмм и посвящено дальнейшее изложение.

Пусть $\{F_0\}$ — набор значений частоты основного тона в сигнале, вычисленных с шагом Δt. Выше отмечалось, что для выделения частоты основного тона использовался метод, описанный в работе [1]. Его результатом для каждого сегмента анализа, кроме частоты основного тона, являлась степень вокализованности участка речи — v_i. Участки сигнала, имеющие степень вокализованности меньше некоторого порога, автоматически отмечались как невокализованные и исключались из дальнейшей обработки. Участки сигнала с частотой основного тона x_i, отмеченные как вокализованные, использовались для построения распределения в виде гистограммы значений с весом v_i. Итоговое распределение, являющееся моделью диктора на этапе обучения, представляло собой гистограмму с $\{h_i\}_{i=0}^{N}$, где h_i — вероятность нахождения частот основного тона заданных пределах фиксированного частотного диапазона.

В качестве способов принятия решения рассматривались следующие методы оценки схожести распределений:

1. Вероятностная модель. Данный метод широко используется при идентификации дикторов и состоит в следующем. Пусть $\{x_i\}_{i=0}^{T}$ — набор значений частоты основного тона в исследуемом сигнале, тогда вероятность принадлежности сигнала диктору описывается выражением:

$$F = \Pi_{i=0}^{T} h_i . \quad (1)$$

2. Евклидово расстояние. Пусть $\{h_{1,i}\}_{i=0}^{N}$ — модель известного диктора, а $\{h_{2,i}\}_{i=0}^{N}$ — модель исследуемой записи, в таком случае степень близости исследуемой записи диктору описывается выражением:

$$F = \Sigma_{i=0}^{N} (h_{1,i} - h_{1,i})^2 . \quad (2)$$

3. Расстояние Кульбака-Лейблера. Пусть $\{h_{1,i}\}_{i=0}^{N}$ — модель известного диктора, а $\{h_{2,i}\}_{i=0}^{N}$ — модель исследуемой записи, в таком случае степень близости исследуемой записи диктору описывается выражением:

$$F = \sum_{i=0}^{N} h_{1,i} h \frac{h_{1,i}}{h_{2,i}} lh \frac{h_{1,i}}{h_{2,i}} . \quad (3)$$

4. Расстояние Хи-квадрат (первый вариант) [3]. Пусть $\{h_{1,i}\}_{i=0}^{N}$ — модель известного диктора, а $\{h_{2,i}\}_{i=0}^{N}$ — модель исследуемой записи, в таком случае степень близости исследуемой записи диктору описывается выражением:

$$F = \sum_{i=0}^{N} \frac{(h_{1,i} - h_{2,i})^2}{h_{1,i}} . \quad (4)$$

5. Расстояние Хи-квадрат (второй вариант) согласно [3]. Пусть $\{h_{1,i}\}_{i=0}^{N}$ — модель известного диктора, а $\{h_{2,i}\}_{i=0}^{N}$ — модель исследуемой записи, в таком случае степень близости исследуемой записи диктору описывается выражением:

$$F = \sum_{i=0}^{N} \frac{(h_{1,i} - h_{2,i})^2}{(h_{1,i} + h_{2,i})} . \quad (5)$$

Так как размер выборки, используемой для построения гистограммы, для различных записей отличался, то дополнительно для каждого метода исследовались три различных метода нормировки. Использовалось

деление полученного значения функции близости на количество частотных диапазонов в модели диктора в модели анализируемой записи или на размер выборки частоты основного тона в идентифицируемой записи.

При получении практических результатов по измерению точности идентификации был проведён ряд экспериментов по вычислению значения эквивалентной ошибки открытой идентификации для различных функций оценки расстояний и нормировки. Используемая для проведения тестирования база содержала записи речи 21 диктора. Для обучения использовался образец речи диктора средней длительностью 145 секунды, для тестирования использовался отличный от обучающего образец речи средней длительностью 127 секунд. Речевые сигналы были записаны из телефонного канала и содержались в аудиофайлах в формате ИКМ с частотой оцифровки 8 кГц. Соотношение сигнал/шум в большей части сигнала составляло не хуже, чем 20 дБ.

Входной сигнал подвергался предобработке, которая удаляла из него участки шума и тишины. Результирующий сигнал сегментировался на блоки в 512 отсчётов с шагом в 256 отсчётов. Для каждого сегмента принималось решение о степени вокализованности, и для вокализованных сегментов вычислялось значение частоты основного тона.

Помимо вычисления эквивалентной ошибки идентификации для анализа результатов использовался метод вычисления ошибки идентификации, отражающей интегральную стоимость решения с субоптимальным порогом выбираемым потребителем — C_{llr} [4].

На рисунке представлены результаты для комбинаций и методов с эквивалентной ошибкой идентификации меньше 25%.

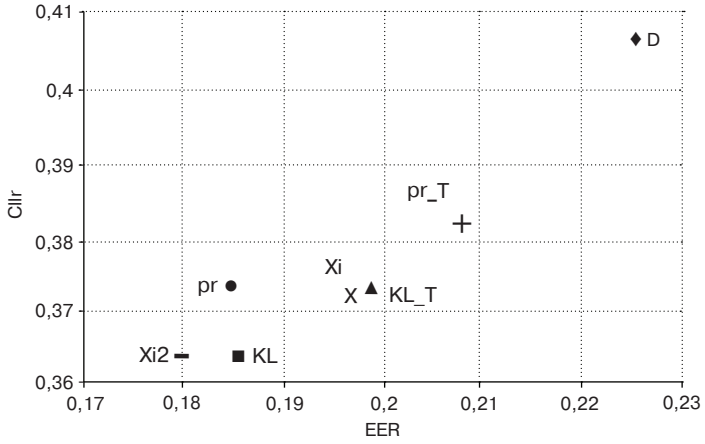

Рис. Соотношение Cllr и EER для различных методов идентификации

Соответствие обозначений исследуемым методам: D — евклидово расстояние, KL — расстояние Кульбака-Лейблера, KL_T — расстояние Кульбака-Лейблера с нормировкой по размеру выборки для идентификации, Xi — расстояние хи-квадрат (первый вариант), Xi2 — расстояние хи-квадрат (второй вариант), pr — вероятностный метод, pr_T — вероятностный метод с нормировкой по размеру выборки для идентификации.

Из полученных данных можно заключить, что вероятностный метод вычисления вероятности принадлежности исследуемой записи к выбранному диктору на основании распределения значений частоты основного тона является приемлемым. Однако

метод хи-квадрат даёт лучшие результаты по обоим способам оценки ошибки идентификации. Нормирование полученных результатов на количество частотных диапазонов в распределении не оказывает улучшения на идентификацию в целом.

В дальнейшем предполагается расширить исследование путём исследования методов сравнения других методов описания распределения плотности вероятности частоты основного тона, а также расширить данный метод за счёт включения в него методов учёта динамических характеристик, связанных с основным тоном.

ЛИТЕРАТУРА

1. *Аграновский А.В., Леднов Д.А., Потапенко А.Н., Репалов С.А., Сулима П.М*. Способ выделения основного тона из речевого сигнала // Патент РФ на изобретение № 2184399 от 22.09.2000, МПК 7 J 10 L 15/00.

2. *Carey M. J., Parris E. S., Lloyd-Thomas H., Bennett S*. Robust Prosodic Features for Speaker Identification // Proc. of ICSLP, 1996, pp.1800–1803.

3. *Боровков А.А*. Математическая статистика: Учебник. 3-е изд. испр. М.: Изд-во физико-математичской литературы, 2007. 704 с. ISBN 9875-94052-141-X.

4. *Niko Brummer, Johan du Preez* «Aplication-Independent Evaluation of Speaker Detection» Computer Speech and Language, 2006. Pp. 230–275.