“Исследование задачи идентификация человека по голосу на основе метода квантования векторов в пространстве цепстральных коэффициентов”

Зинченко Е.Ю., МГУ им. Ломоносова, ф-т ВмиК, каф. АНИ, аспирант.
zineugene@mtu-net.ru

Тезис доклада.

1. Введение.

В докладе развиваются описанные методы идентификации человека (спикера) по записи его речи. Исследуется метод, основанный на использовании априорной информации о голосовых данных отдельного спикера, выделяемой из представительной базы данных записей голоса этого спикера. За характерные параметры голоса выбраны цепстральные коэффициенты[1], вычисляемые на небольших временных интервалах (0.03 секунды) записей голоса. Идентификация основана на применении метода квантования векторов в пространстве вычисленных цепстральных коэффициентов.

Задача является важной в области разработки систем безопасности на охраняемых объектах (банках, военных объектах и.т.д.), при анализе записей черных ящиков, в криминалистике.

Основной акцент в докладе делается на вид структуры кластеров данных в пространстве цепстральных коэффициентов голоса. Выделенные особенности кластеров позволяют улучшить точность идентификации и дать рекомендации для формирования базы данных записей голоса конкретного спикера, которая наиболее полно его характеризует.

Для экспериментов создана база данных записей русской речи спикеров. Записи имеют частоту сэмплирования 8 kHz.

2. Алгоритм идентификации

Алгоритм идентификации человека по голосу состоит из двух этапов: обучения и аккредитации.

Цель этапа *обучения* состоит в выделении в характерных текстонезависимых особенностей голоса q-го спикера на основе априорно заданных записей голоса этого спикера $s^q(t),\quad q=\overline{1,N_q},\quad N_q$-число спикеров, $t=\overline{1,N_t},\quad N_t$-число временных отсчетов в записи.

Этап обучения состоит из следующих шагов:

- Разбиение сигнала $s^q(t)$ на окна (фреймы) $s_k^q(t'),\quad k=\overline{1,N_k},\quad N_k$ - количество окон в сигнале $s^q(t);\ t'=\overline{1,N_t^{win}}$, N_t^{win} - число временных отсчетов в одном фрейме. Чаще всего используют фреймы длиной 0.02-0.04 с [2,3]. Мы использовали фреймы продолжительностью 0.03 секунды с перекрытием в 0.01 секунды.
- Удаление фреймов, содержащих паузы. Отбрасываются те фреймы $s_k^q(t')$, энергия которых
$$E\left(s_k^q(t')\right)=\frac{1}{N_t^{win}}\sum_{\tau=1}^{N_t^{win}}s_k^q(t') \qquad (1),$$
меньше заданного порогового значения δ_E
- Вычисление цепстральных коэффициентов каждого окна
$$f_{s_k^q}(n)=\mathrm{Re}\left[\frac{1}{N_t^{win}}\sum_{l=1}^{N_t^{win}}e^{i\frac{2\pi(l-1)(n-1)}{N_t^{win}}}\ln\left\{\left|\sum_{\tau=1}^{N_t^{win}}s(\tau)e^{-i\frac{2\pi(l-1)(\tau-1)}{N_t^{win}}}\right|\right\}\right],n=\overline{0,N_t^{win}} \qquad (2)$$
$f_{s_k^q}(0)$ соответствует энергии сигнала и не содержит информацию о спикере. Чаще всего для идентификации используют $\vec{f}_k^q=\left(f_{s_k^q}(1),\ldots,f_{s_k^q}(12)\right)$ [1,2,3]. Обозначим набор векторов $\vec{f}_k^q$, соответствующий записи $s^q(t)$, как $F\left(s^q(t)\right)$.
- Вычисление центроидов $\vec{c}_l^q, l=\overline{1,N_c}$ кластеров множества векторов $\left\{\vec{f}_k^q\right\}$ методом К-средних.

Каждый набор векторов $C_q=\left\{\vec{c}_l^q, l=\overline{1,N_c}\right\}$ (словарь *q*-го спикера) характеризует q-го спикера.

На этапе *аккредитации* осуществляется идентификация спикера в заданной записи $x(t)$. Идентификация осуществляется на основе близости цепстральных коэффициентов $F(x(t))$ к словарям спикеров C_q. В качестве меры близости записи $x(t)$ к словарю C_q мы будем использовать следующий функционал:

$$d\left(F\left(x(t)\right),C_q\right)=\frac{1}{N_k}\sum_{k'=1}^{N_k}\min_{l'=1,\ldots,N_c}\left\|\vec{f}_{k'}-\vec{c}_{l'}^{\,q}\right\| \tag{3}$$

Тогда $$q=\arg\min_{q'=1,\ldots,N_q}\left\{d\left(F\left(x(t)\right),C_{q'}\right)\right\} \tag{4}$$

индекс идентифицированного спикера.

3. Анализ качества кластеризации.

Ясно, что качество идентификации находится в прямой зависимости от структуры кластеров в пространстве цепстральных коэффициентов. Для оценки качества кластеризации во многих работах [4,5] используется параметр F_{mecm}, который есть отношение дисперсии между центроидами к средней дисперсии элементов внутри кластеров. Чем больше значение этого параметра, тем кластеризация лучше.

$$F_{mecm}=\frac{\sum_{k=1}^{K}\left\|\frac{1}{L}\sum_{l'=1}^{L}c_{l'}-\vec{c}_{l'}\right\|^2}{\sum_{l'=1}^{N_c}\frac{1}{n_{l'}}\sum_{k':g_k=l'}d_E\left(\vec{f}_{k'},c_{l'}\right)^2} \tag{5},$$

где n_l количество элементов в *l*-ом кластере, g_k номер кластера, которому соответствует $\vec{f}_k$.

На рис. 1. показана зависимость F_{mecm} от числа кластеров N_c, на которые разбиваем множество. На графике видно, что большое значение параметра достигается в точке $N_c=8$. В точках левее $N_c=64$ скорость изменения параметра уменьшается. При кластеризации на $N_c=64$ число кластеров, система показала 100 % идентификации в лабораторных условиях. Это совпадает с результатами, полученными в других работах. В работе [2] так же указывается, что точность идентификации 100 % достигается при $N_c=64$. Однако значение параметра F_{mecm} много меньше 1. Это говорит о сильной смешанности данных.

На рис. 2.(а, б) показана двумерная проекция распределения множества $F\left(x(t)\right)$ для двух разных спикеров. Стрелки 1,2 указывают на разные кластеры. Стрелка 3 указывает на общий кластер. Эта ситуация характерна для речевых данных. Можно предполагать, что существуют кластеры , в которых содержится информация о спикере и существуют кластеры, в которых эта информация подавлена информацией другого типа, например информацией содержании произносимой фразы. Из рис. 2 также видно, что форма кластеров имеет эллиптический виде. Это говорит о том, что большую точность можно достичь, если проводить идентификацию на основе разделяющих функций.

4. Выводы.

В докладе исследован метод идентификации человека по голосу, на основе метода квантования векторов в пространстве цепстральных коэффициентов. Показано, что для высокой точности идентификации, кластеризацию надо проводить для большого числа кластеров (64).

Исследована внутренняя структура кластеров. Показано, что внутрикластерная дисперсия выше, чем межкластерная дисперсия. Выявлено, что существуют кластеры, уникальные для каждого спикера, и существуют кластеры, которые у разных спикеров одинаковые. Показано, что форма кластеров эллиптична. Поэтому перспективными выглядят подходя, связанные с применением методов идентификации, основанных на разделяющей функции.

Литература.

[1] Furui, S.: Comparison of Speaker Recognition Methods Using Statistical Features and Dynamic Features, IEEE Transactions on Acoustics, Speech, and Signal Processing, No.3, June 1981

[2] Kinnunen T., Kilpelainen T., Franti P. Comparison of clustering algorithms in speaker identification., *IASTED Int. Conf. on Signal Processing and Communications (SPC'00)*, Marbella, Spain, 222-227, 2000.

[3] T. S. Chang, S.D. Van Hooser. Two new methods for speaker recognition using cepstral analysis., CNS-Spring, 1996.

[4] Kinnunen T., Franti P.: Is speech data clustered? – statistical analysis of cepstral features., *European Conf. on Speech Communiation and Technology, (EUROSPEECH'2001)*, Aalborg, Denmark, Vol. 4, pp. 2627-2630, September, 2001.

[5] Özgür Devrim Orman. Frequency analysis of speaker identification performance. Master of Science Thesis, Bogazici University, 2000.

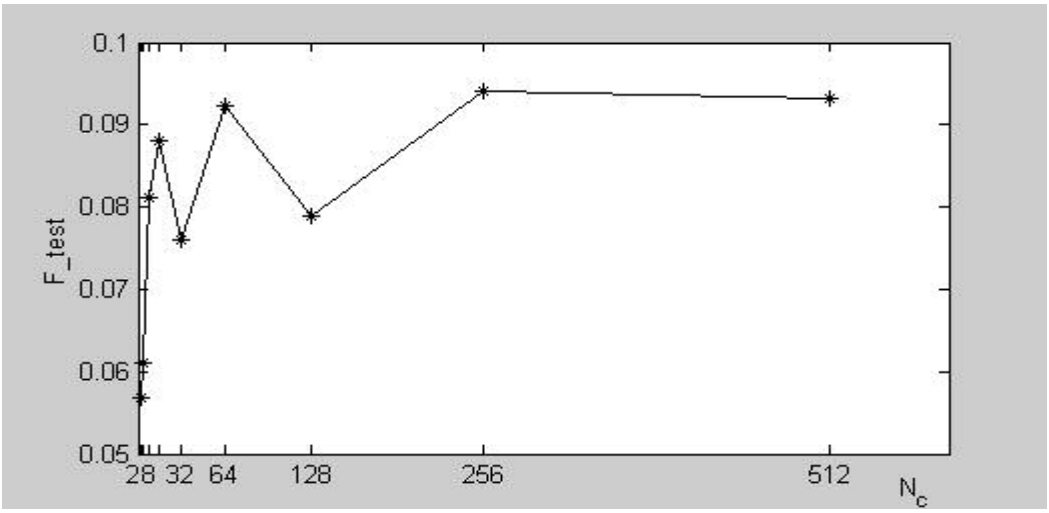
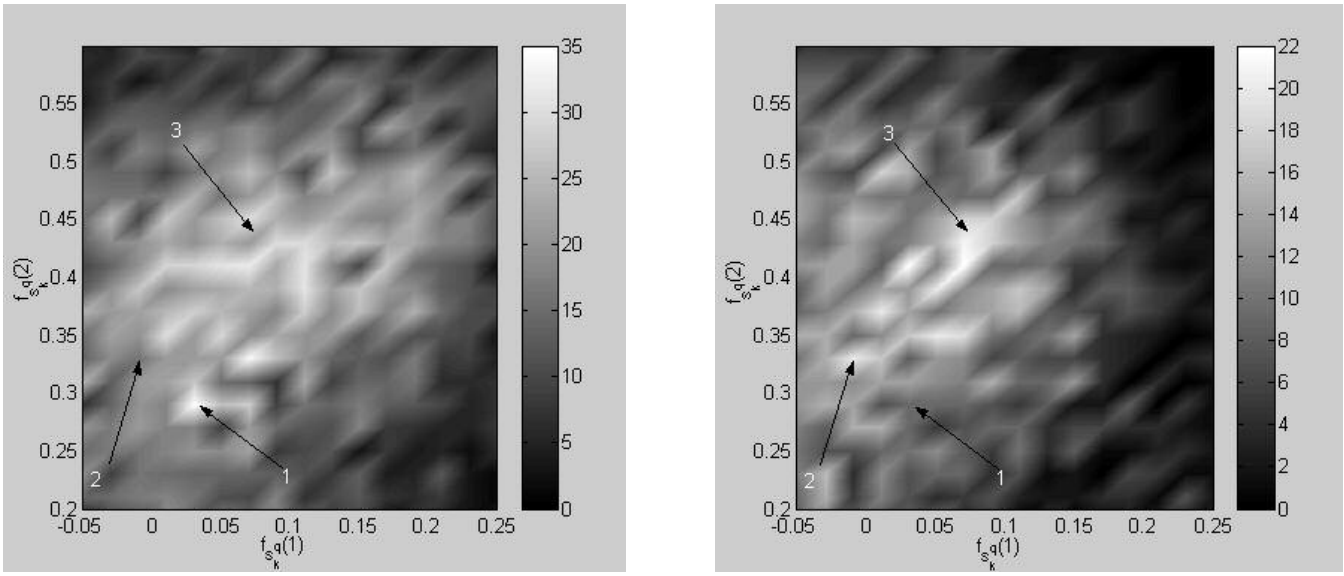


Рис. 1. Зависимость параметров F_{mecm} от числа кластеров, на которые разбиваем множество.



а.) б.)

Рис. 2 (а, б) Проекция распределения первых двух цепстральных коэффициентов разных спикеров: (а) – первый спикер. (б) – пятый спикер. Стрелки 1,2 указывают на разные кластеры. Стрелка 3 указывает на общий кластер.