УДК 004.89

ПОРІВНЯЛЬНИЙ АНАЛІЗ ЕФЕКТИВНОСТІ ІДЕНТИФІКАЦІЇ ДИКТОРІВ ЗА СПЕКТРАЛЬНИМИ ХАРАКТЕРИСТИКАМИ

Сілаічев М.В.
e-mail: mykhailo.silaichev@nure.ua
Науковий керівник – проф. Безрук В.М.
Харківський національний університет радіоелектроніки, каф. ІМІ
м. Харків, Україна

This work investigates the accuracy of automatic speaker identification using three spectral features: Mel-frequency cepstral coefficients (MFCC), spectral centroid, and spectral roll-off. An experiment was conducted on eight speakers based on a Java-implemented system. Mathematical descriptions of the features and similarity metrics used in the software implementation are provided. The results can be used in the design of voice biometric systems where a balance between computational complexity and reliability is required.

Біометрична ідентифікація людини за голосом залишається актуальним завданням у системах безпеки, голосових помічниках і криміналістиці. Найпоширенішим підходом є виділення з мовного сигналу компактних ознак, що характеризують тембральні та динамічні особливості голосу, з подальшим порівнянням з еталонними шаблонами (голосовими відбитками). Серед багатьох можливих дескрипторів особливе місце посідають мел-частотні кепстральні коефіцієнти (MFCC), які зарекомендували себе як ефективний інструмент. Однак на практиці часто використовуються й простіші спектральні характеристики, такі як центроїд і рол-оф. Мета даної роботи – кількісно порівняти результативність ідентифікації дикторів при використанні різних спектральних ознак на фіксованій базі голосових записів реалізацій для різних дикторів із використанням методу статистичних випробовувань.

У розробленій програмній системі на мові Java спектральні ознаки дикторів обчислюються для кожної реалізації. Мовні сигнали діляться на перекривні кадри довжиною 1024 відліки з кроком 256 відліків (частота дискретизації 22050 Гц). До кожного кадру застосовується вікно Ханна. Далі для кадра розраховується спектр потужності за допомогою швидкого перетворення Фур'є (ШПФ). На основі спектра потужності вилучаються три типи характеристик: мел-частотні, спектральний центроїд та спектральний рол-оф.

Мел-частотні кепстральні коефіцієнти (MFCC) обчислюються за стандартною методикою. Застосування вагової функції Ханна до кадру:

$$\omega(n) = 0.5\left(1 - cos\ cos\ \frac{2\pi n}{N-1}\right), n = 0, \dots, N-1, N = 1024.$$

Розрахунок амплітудного спектра X(k) за допомогою ШПФ та отримання спектра потужності $P(k) = |X(k)|^2$. Застосування набору трикутних смугових фільтрів у шкалі mel: $mel(f) = 2595(1 + f/700)$. Енергія на виході кожного фільтра E_j логарифмується.

Дискретне косинусне перетворення (ДКП) для отримання коефіцієнтів:

$$MFCC_i = \sum_{j=1}^{M} \quad log\ log\ (E_j)\ cos\ cos\ \left(i\frac{\pi}{M}(j-0.5)\right)\ , i = 1, \dots, 13,$$

де ММ – кількість фільтрів (у системі – 40). Для кожного кадру виходить вектор MFCC розмірності 13. Часова послідовність таких векторів усереднюється, формуючи єдиний вектор ознак диктора a_{MFCC}.

Спектральний центроїд характеризує «центр мас» спектра. Для кадру він обчислюється за формулою:

$$C = \frac{\sum_{k=0}^{K-1} f(k)P(k)}{\sum_{k=0}^{K-1} P(k)},$$

де $K = N/2 = 512$ – кількість бінів спектра потужності, $f(k) = k \times F_s/(2K)$ – частота, що відповідає біну k (у Гц), $F_s = 22050$ Гц – частота дискретизації. Підсумковою ознакою диктора є середнє значення центроїда по всіх кадрах: $\underline{C}$.

Спектральний рол-оф визначає частоту, нижче якої зосереджена задана частка (у нашому випадку 85%) загальної енергії спектра:

$$R = f(m), \qquad \text{де } m = min\left\{\sum_{k=0}^{m} P(k) \geq 0.85 \sum_{k=0}^{K-1} P(k))\right\}$$

Тут $f(m)$ – частота біна m. Середнє по кадрах значення $\underline{R}$ використовується як характеристика диктора.

В експерименті брали участь записи восьми дикторів. Для кожного диктора було два фрагменти мовного сигналу довжиною одну хвилину: один використовувався для створення еталонного голосового відбитка (навчання), другий – для тестування (ідентифікація). Усі записи є монофайлами формату AU з частотою дискретизації 22050 Гц.

Для кожного диктора на основі навчального файлу обчислювалися середні вектори/значення ознак: a_{MFCC} (середній вектор MFCC), $\underline{C}$ (середній спектральний центроїд), $\underline{R}$ (середній спектральний рол-оф). При тестуванні для невідомого запису обчислювалися аналогічні усереднені характеристики $b_{MFCC}, \underline{C_t}, \underline{R_t}$.

Для MFCC використовувалася косинусна подібність:

$$S_{MFCC} = \frac{a_{FMCC} \times b_{FMCC}}{\|a_{FMCC}\| \times \|b_{FMCC}\|}.$$

Для центроїда та рола-офа застосовувалася експоненційна функція від абсолютної різниці:

$$s_C = exp\ exp\left(-\frac{\left(\underline{C}-\underline{C_t}\right)^2}{2\sigma_C^2}\right),\ \ S_R = exp\ exp\left(-\frac{\left(\underline{R}-R_t\right)^2}{2\sigma_C^2}\right),$$

де параметри підібрані емпірично:$\sigma_C = 500$ Гц, $\sigma_R = 1000$ Гц.

Ідентифікація проводилася за принципом найближчого сусіда: невідомий диктор відносився до того еталона, для якого подібність максимальна і перевищує поріг $\mathrm{T} = 0.75$.

За результатами досліджень найкращий результат показали MFCC, що узгоджується з відомими дослідженнями: MFCC компактно описують спектральну обвідну, враховуючи нелінійне сприйняття частоти людиною. Завдяки використанню 13 коефіцієнтів вони захоплюють як формантну структуру, так і загальний тембр голосу.

Спектральний центроїд виявився менш інформативним. Це пояснюється тим, що центроїд сильно залежить від гучності та частотного діапазону запису, які можуть варіюватися навіть в одного диктора за різних умов. Крім того, центроїд – це лише одне число, яке втрачає деталі спектра.

Найнижчу точність дав спектральний рол-оф. Він характеризує лише розподіл енергії в низькочастотній області, що часто є подібним у різних людей. Помилки виникали здебільшого через близькість значень рола-офа у дикторів однієї статі.

Отримані результати можуть бути використані при проєктуванні голосових біометричних систем, де потрібен баланс між обчислювальною складністю та надійністю. Подальші дослідження передбачають включення динамічних характеристик (дельта- та дельта-дельта-коефіцієнтів) і застосування методів машинного навчання.

Список використаних джерел:

1. Davis S., Mermelstein P. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. IEEE Transactions on Acoustics, Speech, and Signal Processing. 1980. Vol. 28, no. 4. P. 357–366.

2. Six J., Cornelis O., Leman M. TarsosDSP, a Real-Time Audio Processing Framework in Java. Proceedings of the 53rd AES Conference (AES 53rd). 2014. 6 p.

3. Peeters G. A large set of audio features for sound description (similarity and classification) in the CUIDADO project. Technical report, IRCAM. 2004. 25 p.