УДК 681.58

АНАЛІЗ ВПЛИВУ ПАРАМЕТРІВ ОБРОБКИ ЗВУКОВОГО СИГНАЛУ НА ЯКІСТЬ РОЗПІЗНАВАННЯ ГОЛОСОВИХ КОМАНД

Дюжаєв Л. П.[1], к.т.н., доцент; Коваль В. Ю.[2], магістрант
[1] Національний технічний університет України
«Київський політехнічний інститут», м. Київ, Україна
[2] ТОВ «Центральна промислова група», м. Київ, Україна

ANALYSIS OF THE INFLUENCE OF SOUND SIGNAL PROCESSING PARAMETERS ON THE QUALITY VOICE COMMAND RECOGNITION

Dyuzhayev L. P.[1], PhD, Associate Professor, Koval V. Yu.[2], Undergraduate Student,
[1] National Technical University of Ukraine «Kyiv Polytechnic Institute», Kyiv, Ukraine
[2] LLC «Central Industrial Group», Kyiv, Ukraine

Вступ

На сьогоднішній день великого поширення набули різноманітні інформаційно-керуючі системи. Такі системи особливо зручні, коли оператор може спілкуватися з ними природним для себе чином — за допомогою голосових команд (ГК). Тому велика увага дослідників приділяється створенню голосових інформаційно-керуючих систем (ГІКС). Такі системи особливо корисні при керуванні транспортом та на виробництві, коли необхідно керувати виконавчими механізмами, отримувати інформацію про їх стан і стан навколишнього оточення і таким чином знизити навантаження на оператора.

Для задачі голосового керування різними пристроями необхідне розпізнавання окремих (ізольованих) голосових команд. Як правило, такий спосіб управління вимагає високої надійності (не менше 95% точності розпізнавання голосових команд), при цьому слід врахувати, що часто команди вимовляються в умовах підвищеної зашумленості.

Всі відомі на сьогодні методи і алгоритми в розпізнаванні мови не дають можливості явним чином визначити які параметри голосового сигналу можуть дати найкращі результати.

Постановка задачі

В даній статті розглянуто етапи первинної обробки аудіо сигналів, алгоритм отримання акустичних ознак голосової команди, реалізується та проводиться моделювання системи розпізнавання голосових команд алгоритмом динамічного викривлення часу (ДВЧ), досліджується вплив на результат розпізнавання голосової команди таких параметрів обробки звукового сигналу:

- частота дискретизації;

- тривалість фрейму;
- кількість відліків перетворення Фур'є;
- вид віконної функції;

Теоретичні викладки

Сучасні системи розпізнавання мови як правило мають ієрархічну модульну структуру. На першому рівні виконується попередня обробка та виділення акустичних ознак, які характеризують голосову команду. Одним з найуживаніших на сьогодні методів — є виділення мел-частотних кепстральних коефіцієнтів (*Mel-Frequency Cepstral Coefficients* або *MFCC*). Мел — психофізична одиниця висоти звуку, що пов'язана з частотою за формулою (3) [1,2]. Отримані на основі цього методу ознаки володіють рядом корисних властивостей — вони легко розраховуються, дають компактне представлення голосової команди, стійкі до шумових завад з навколишнього середовища.

Наступний рівень систем розпізнавання голосових команд — лінгвістичний. В нього входить процедура пошуку вимовленої команди по словникам еталонів. При розпізнаванні окремих голосових команд, диктор вимовляє слово без оточуючого контексту. Навчання таких систем є трудомісткою задачею і для підвищення надійності зазвичай використовують великі навчальні вибірки (від 5 та більше варіантів вимови однієї голосової команди). Кожна команда записується в словник еталонів як набір мел-частотних кепстральних коефіцієнтів. Типова структура такої системи наведена на рис. 1.

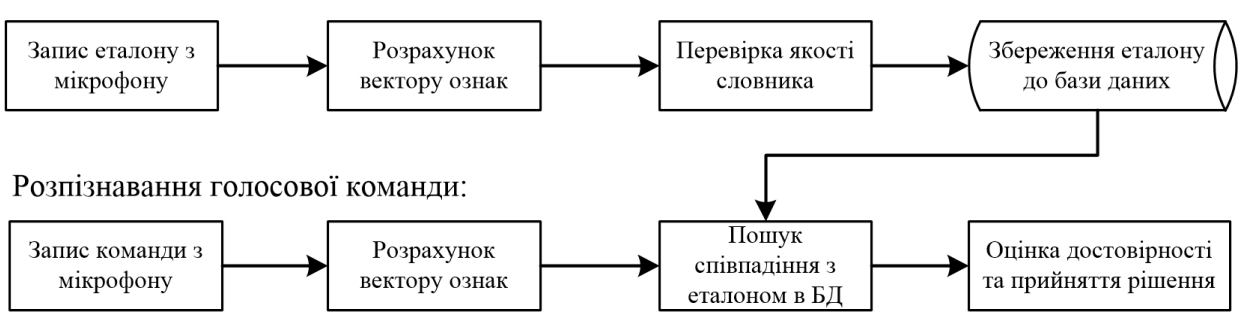

Рис. 1 Структура системи розпізнавання голосових команд

Алгоритм розрахунку мел-частотних кепстральних коефіцієнтів

Такий метод отримання ознак є одним з найпоширеніших як в системах розпізнавання дикторів так і в системах розпізнавання мови.

В вхід алгоритму подається послідовність відліків ділянки сигналу, що досліджується на даній ітерації, $q_0,\ldots q_{N-1}$. На цю послідовність накладається вагова функція і після цього виконується дискретне перетворення Фур'є. Вагова функція використовується для зменшення спотворень в аналізі Фур'є, які викликані скінченністю вибірки. На практиці в якості вагової функції часто використовуються вікно Хеммінга (1) та вікно Ханна (2).

$$w_n = 0.54 - 0.46 \cdot \cos\left(2\pi \frac{n}{N-1}\right), n = 0,....N-1 \quad (1)$$

$$w_n = 0.5 \cdot \left(1 - \cos\left(2\pi \frac{n}{N-1}\right)\right), n = 0,....N-1 \quad (2)$$

де N — довжина вікна виражена у відліках.

Тоді дискретне перетворення Фур'є зваженого сигналу можна записати у наступному вигляді:

$$X_k = \sum_{n=0}^{N-1} x_n w_n e^{\frac{-2\pi i}{N} kn}, k = 0,..., N-1$$

Значення індексів *k* відповідає наступним частотам:

$$f_k = \frac{F_s}{N} k, k = 0,...N/2,$$

де F_s — частота дискретизації сигналу.

Отримане представлення сигналу у частотній області розбивають на діапазони за допомогою банка трикутних фільтрів. Межі фільтрів розраховуються в шкалі мел. Дана шкала є результатом досліджень здібностей людського вуха до сприйняття звуків на різних частотах. Перехід в мелчастотну [3] область здійснюють за наступною формулою:

$$M(f) = 1127 \cdot \ln\left(1 + \frac{f}{700}\right) \quad (3)$$

Нехай N_{FB} — кількість фільтрів, (f_H, f_B) — досліджуваний діапазон частот. Тоді цей діапазон переводять в шкалу мел, розбивають на N_{FB} рівномірно розподілених діапазони та розраховують відповідні межі в області лінійних частот. Позначимо через $H_{m,k}$ — вагові коефіцієнти отриманих фільтрів. Фільтри застосовуються до квадратів модулів коефіцієнтів перетворення Фур'є. Отримані значення логарифмуються:

$$e_m = \ln\left(\sum_{k=0}^{N} |X_k|^2 H_{m,k}\right), m = 0,...N_{FB} - 1$$

Заключним етапом в розрахунку MFCC коефіцієнтів є дискретне косинусне перетворення:

$$C_i = \sum_{m=0}^{N_{FB}-1} e_m \cos\left(\frac{\pi i(m+0.5)}{N_{FB}}\right), i = 1,...N_{MFCC}$$

На практиці кількість коефіцієнтів N_{MFCC} дорівнює 12 (окрім першого), оскільки вони містять 95% корисної інформації про звуковий сигнал.

Рис. 2 Алгоритм отримання MFCC коефіцієнтів

Метод динамічного викривлення часу

В задачах розпізнавання мови цей метод використовується для оцінки міри схожості між вхідною голосовою командою та еталоном з бази даних. Цей метод дозволяє порівнювати різні в тривалості зразки, тобто розпізнавання команди не залежить від темпу її вимови.

Нехай порівнюється два зразки голосових команд, які представлені у вигляді масиву векторів, в нашому випадку це набір MFCC коефіцієнтів:

$$S = \{\bar{s}_0, \bar{s}_1, ...\bar{s}_n\} \text{ та } Q = \{\bar{q}_0, \bar{q}_1, ...\bar{q}_m\}$$

Різниця між векторами двох зразків визначається послідовністю станів C_k та позначається:

$$F() = C_0, C_1, ...C_k ..., C_K,$$

де C_0 та C_K – початкові та кінцеві стани, *F()* – функція часового вирівнювання, що відображає накладання часової області одного зразка на часову область другого зразка.

Метод ДВЧ полягає в тому, що відбувається пошук такої функції $F()$, при якій шлях зі стану C_0 в стан C_K буде оптимальним, тобто буде отримано мінімальну відстань між двома зразками.

При побудові оптимального шляху, на кожному кроці алгоритму, використовується основна формула ДВЧ [4]:

$$d_{i,j} = \min \begin{Bmatrix} d_{i,j-1} + r(\bar{s}_i, \bar{q}_j) \\ d_{i-1,j-1} + r(\bar{s}_i, \bar{q}_j) \\ d_{i-1,j} + r(\bar{s}_i, \bar{q}_j) \end{Bmatrix}, \quad (4)$$

де *i=0…N, j=0…M*

В якості відстані між двома векторами використовується евклідова метрика:

$$r(\bar{s}_i, \bar{q}_j) = \sum_{k=0}^{V} (s_k - q_k)^2$$

, де V — розмірність векторів ознак

Псевдо код алгоритму наведено нижче [5]:

Вхід: *S: вектор довжиною n, Q: вектор довжиною m.*

Вихід: *Міра близькості DTW.*

1: *Ініціалізація D(i,1) ← для кожного i.*

2: *Ініціалізація D(1,j) ← для кожного j.*

3: **цикл від** $i=1$ **до** n

4. **якщо** $2 \le i \le n$ **то**

5: **цикл від** $j=1$ **до** m

6: **якщо** $2 \le j \le m$ **то**

7: *Використати умову (4) для визначення* $d_{i,j}$

8: **кінець циклу**

9: **кінець циклу**

10: **повернути** $d_{n,m}$

На виході процедури порівняння отримується деяке число (міра близькості), що представляє собою величину, зворотну ступеню схожості між порівнюваними сигналами.

Вплив параметрів обробки звукового сигналу на якість розпізнавання голосових команд

Розглянемо вплив таких параметрів обробки звукового сигналу як частота дискретизації, тривалість фрейму, на які розбивається звуковий сигнал, віконна функція кількість вибірок для перетворення Фур'є. Система моделюється в середовищі MatLab. Розпізнавання проводилось по словнику з 50 команд, вимовлених одним диктором.

Для задач розпізнавання мови прийнято використовувати частоту дискретизації від 8 кГц, оскільки діапазон частот людського голосу лежить в межах 300-4000 Гц і згідно теореми Котельнікова частота дискретизації має

бути вдвічі вищою за найбільшу частоту в оброблюваному сигналі. Зменшення частоти дискретизації призводить до збільшення впливу шумів на розпізнавання команд. Підвищення частоти дискретизації підвищує точність розпізнавання, але значно збільшується час обробки звукових даних.

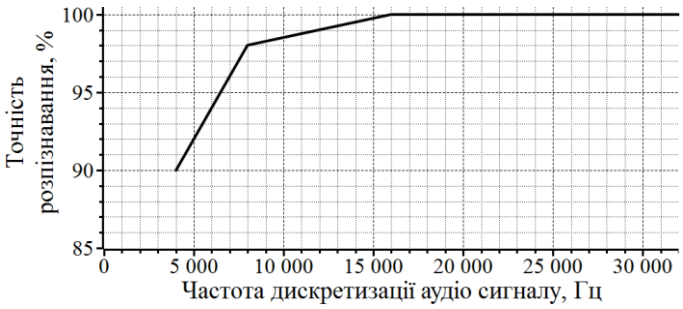

Рис. 3 Залежність точності розпізнавання від частоти дискретизації аудіо сигналу

Залежність якості розпізнавання від частоти дискретизації можна побачити на рис. 3.

Оскільки мовний сигнал являє собою змінний в часі процес, то його спектральний опис базується на концепції коротко часового аналізу [6]. Для цього звуковий сигнал розбивається на рівні відрізки,

що перекриваються, які називаються фреймами або кадрами, в межах яких властивості сигналу мало-змінні і його можна вважати квазістаціонарним. Зазвичай тривалість фрейму обирають рівною 10-100 мс. На кожен такий фрейм накладається віконна функція. Після цього для кожного фрейму виконується спектральний аналіз, в результаті отримується послідовність спектрів. Ця послідовність спектрів, що описує звуковий сигнал зазвичай називають динамічною спектрограмою.

Висновки

Проведені дослідження системи розпізнавання голосових команд методом динамічного викривлення часу показують, що в задачах такого типу найкращий результат можна отримати при розбитті звукового файлу на фрейми тривалістю від 70 до 120 мс, а для зважування використовувати вікно Хеммінга. Найменша частота дискретизації, що забезпечує достатню точність розпізнавання складає 8 кГц, підвищення частоти дискретизації до 16 кГц збільшує точність, але в цей же час збільшується час обробки аудіо даних, тому можна вважати оптимальним значенням частоти дискретизації з точки зору відношення якість/швидкодія є 8 кГц. Для забезпечення надійного розпізнавання команд кількість вибірок перетворення Фур'є має складати щонайменше 512.

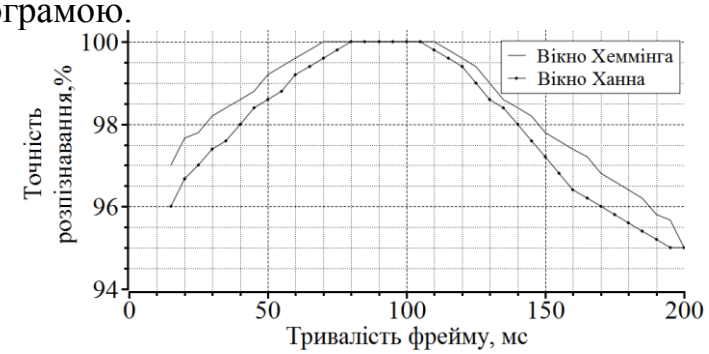

Рис. 4 Залежність точності розпізнавання від тривалості фрейму

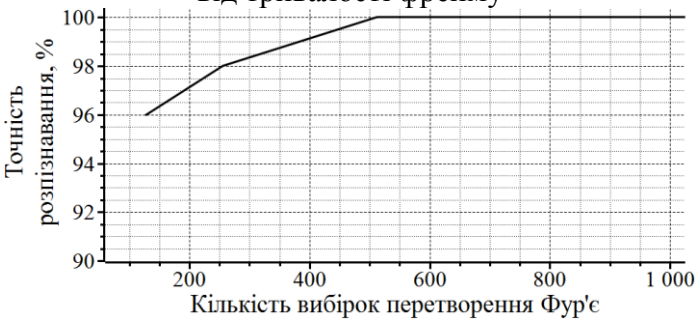

Рис. 5 Залежність точності розпізнавання від кількості вибірок перетворення Фур'є

Визначені параметри в подальшому заплановано застосовувати для створення вбудованих систем розпізнавання голосових команд з малим споживанням енергії.

Перелік посилань

1. Яцковський В. С. Алгоритм оцінювання темпу музикальних сигналів / В. С. Яцковський, Д. М. Бруслік // Електроніка та системи управління. — 2012. — № 31. — С. 5—9.

2. Dhingra S. D. Isolated speech recognition using MFCC and DTW / S. D. Dhingra, G. Nijhawan, P. Pandit // International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering. — 2007. — Vol. 2, No 8. — pp. 4085—4092.

3. Гладышев К. К. Информативные признаки на основе линейных спектральных корней в системах распознавания команд: автореф. дис. на соискание ученой степени канд. техн. наук: спец. 05.13.01 – «Системный анализ, управление и обработка инфор-

мации» / Гладышев Константин Константинович; С–Пб. госуд. унив. телекоммуникаций им. проф. М. А. Бонч–Бруевича. — СПб, 2010. — 16 с.

4. Al–Naymat G. SparseDTW: A Novel Approach to Speed up Dynamic Time Warping. / G. Al–Naymat, S. Chawla, J. Taheri // The 2009 Australasian Data Mining. — 2009. — Vol. 101 — pp. 117—127.

5. Muda L. Voice Recognition Algorithms using Mel–Frequency Cepstral Coefficient (MFCC) and Dynamic Time Warping (DTW) Techniques. / L. Muda, M. Begam, I. Elamvazuthi // Journal of computing. — 2010. — Vol. 2, No 3.

6. Колоколов А. С. Обработка сигнала в частотной области при распознавании речи. / А. С. Колоколов // Проблемы управления. — № 3. — 2006 г. — С. 13—18.

References

1. Yatskovsky V. S. and Bruslik D. N. (2012) Algorithm of tempo estimation of musical signals. *Electronics and Control Systems*. No 31, pp. 5-9.

2. Dhingra S. D. and Nijhawan G. (2007) Isolated speech recognition using MFCC and DTW. *International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering*. Vol. 2, No. 8, pp. 4085 – 4092.

3. Gladyshev K. K. (2010) *Informativnye priznaki na osnove linejnyh spektral'nyh kornej v sistemah raspoznavanija komand*. Diss. Cand.. Tekhn. nauk [Informative features based on linear spectral roots in commands recognition systems. Cand. Sci. diss.]. Saint-Petersburg, The Bonch-Bruevich Saint - Petersburg State University of Telecommunications, 16 p. Available at: www.sut.ru/doci/nauka/avtoref/glad.doc

4. Al-Naymat G., Chawla S. and Taheri J. (2009) SparseDTW: A Novel Approach to Speed up Dynamic Time Warping. *The 2009 Australasian Data Mining*. Vol. 101, Melbourne, Australia, ACM Digital Library, pp. 117-127.

5. Muda L., Begam M. and Elamvazuthi I. (2010) Voice Recognition Algorithms using Mel–Frequency Cepstral Coefficient (MFCC) and Dynamic Time Warping (DTW) Techniques. *Journal of computing*. Vol. 2, No 3, pp. 138–143.

6. Kolokolov A.S. (2006) Frequency domain signal processing in speech recognition. *Control Science*. No 3, pp. 13-18. (In Russian)

Дюжаєв Л .П., Коваль В. Ю. ***Аналіз впливу параметрів обробки звукового сигналу на якість розпізнавання голосових команд.*** *В роботі розглянуто структуру системи розпізнавання голосових команд, алгоритм виділення мел-кепстральних коефіцієнтів та їх порівняння методом динамічного викривлення часу. В системі зі словником з п'ятдесяти команд вимовлених одним диктором було досліджено вплив на якість розпізнавання голосової команди таких параметрів як: частоти дискретизації, тривалості фрейму, кількості вибірок Фур'є, виду віконної функції на якість розпізнавання голосової команди.*

Ключові слова*: розпізнавання мови, голосові команди, мел-кепстральні коефіцієнти, динамічне викривлення часу.*

Дюжаев Л. П., Коваль В. Ю. ***Анализ влияния параметров обработки звукового сигнала на качество распознавания голосовых команд.*** *В работе рассмотрено структуру системы распознавания голосовых команд, алгоритм выделения мел-кепстральных коэффициентов и их сравнение методом динамического искажения времени. В системе со словарем из пятидесяти команд произнесенных одним диктором было исследовано влияние на качество распознавания голосовых команд таких параметров как: частота дискретизации, продолжительность фрейма, количество выборок Фурье, вид оконной функции.*

Ключевые слова: *распознавания речи, голосовые команды, мел-кепстральные коэффициенты, динамическое искажение времени.*

Dyuzhayev L. P., Koval V. Yu. ***Analysis of the influence of sound signal processing parameters on the quality voice command recognition.***

Introduction. Recognition of single (isolated) voice commands for the task of voice control over different devices is required. Typically, this control method requires high reliability (at least 95% accuracy voice recognition). It should be noted that voice commands are often pronounced in high noisiness. All presently known methods and algorithms of speech recognition do not allow clearly to determine which parameters of sound signal can provide the best results.

The main part. On the first level of voice recognition (preprocessing and extracting of acoustic features that have a number of useful features) they are easily calculated, providing a compact representation of the voice commands that are resistant to noise interference. On the next level given command is looked for in the reference dictionary. Input file has to be divided into frames to get MFCC coefficients. Each frame is measured by a window function and processed by discrete Fourier transform. The resulting representation of signal in the frequency domain is divided into ranges using a set of triangular filters. The last step is to perform discrete cosine transform. Method of dynamic time warping allows to get a value, inverse of degree of similarity between given command and a reference.

Conclusions. Research has shown that in the field of voice commands recognition optimum results in terms of quality / performance can be achieved using the following parameters of sound signal processing:8 kHz sample rate, frame duration 70-120 ms, Hamming weighting function of a window, number of Fourier samples is 512.

Keywords: *speech recognition, voice commands, mel-cepstral coefficients, dynamic time warping.*