УДК 681.327.12

М.М.Биков, к.т.н., проф.; В.В.Ковтун, к.т.н., доц.; А. Раїмі, PhD, доц.

МЕТОД НОРМАЛІЗАЦІЇ ТРИВАЛОСТІ ЗВУЧАННЯ ПАРОЛЬНИХ ФРАЗ ДЛЯ СИСТЕМИ РОЗПІЗНАВАННЯ МОВЦІВ

Авторами запропоновано метод нормалізації тривалості звучання записів мовних сигналів, орієнтовану на інтеграцію в систему ідентифікації мовця за індивідуальними особливостями його голосу. Метод дозволяє мінімізувати втрати інформації про індивідуальні властивості мовних сигналів при зміні тривалості їх звучання за рахунок аналізу та сегментації мовного сигналу відповідно до його вмісту та застосування динамічного коефіцієнту стиснення, який визначається результатами сегментації.

Ключові слова: ідентифікація мовця, розпізнавання голосів, сегментація мовних сигналів, нормалізація тривалості звучання.

Втуп

Методи цифрової обробки і розпізнавання мовних сигналів в даний час інтенсивно розвиваються. Це перш за все обумовлено прогресом в області цифрової мікросхемотехніки, завдяки якому з'явилася реальна можливість виготовлення складної цифрової апаратури передачі повідомлень, а також цифрових пристроїв розпізнавання мови, синтезу мови. Перші зразки таких пристроїв, вже освоєні промисловістю, викликали підвищений інтерес розробників до можливостей і залучили нових прихильників цього напрямку досліджень до вивчення існуючих та розробки нових методів і алгоритмів цифрової обробки мови [1].

Розпізнавання голосу мовця засновано на аналізі унікальних характеристик мови, обумовлених анатомічними особливостями (форма артикуляторного тракту, лінійні параметри голосових низок) та набутими звичками (гучність, манера, швидкість мови). Перевагою даного підходу до розпізнавання особи є, насамперед, невіддільність мовної інформації від об’єкта, природність мовного спілкування для людини.

Однією з основних проблем при розпізнаванні голосу є те, що парольна фраза або слово може бути вимовлено з довільною швидкістю. Через це тривалість вимовленого слова не співпадає з тривалістю відповідного еталона [2], що складає актуальну проблему при розробці систем розпізнавання мовців. Більшість відомих напрацювань передбачають рівномірну модифікацію тривалості мовних записів без урахування особливостей темпоральних змін окремих класів звуків, що призводить до втрати індивідуальних особливостей мовлення, наприклад, такої нормативної ознаки, як частота основного тону, яка спостерігається в діапазоні низьких частот, яки суттєво спотворюються звичайними методами нормалізації тривалості звучання.

Кореляційний метод виділення частоти основного тону в системах розпізнавання мовців

Метод оснований на таких математичних засадах. Нехай $s(n) = s(0), s(1), ..., s(N-1)$ - стаціонарний часовий ряд з нульовим середнім значенням. Відповідно з теоремою Уолда [1], коваріаційну функцію такого процесу можна представити у вигляді

$$\gamma(k) = \int_{-\pi}^{\pi} \cos(k\omega) dG(\omega), \qquad (1)$$

де $G(\omega)$ - спектральна щільність потужності сигналу, $k = 0, \pm 1, \pm 2, \ldots$ - часове зміщення.

Автокореляційна функція – нормована величина $R(k) = \gamma(k)/\gamma(0)$, яку з врахуванням (1) можна описати таким рівнянням

$$R(k) = \frac{\int_{-\pi}^{\pi} \cos(k\omega) dG(\omega)}{\int_{-\pi}^{\pi} dG(\omega)}, \quad k = 0, \pm 1, \pm 2, \ldots \qquad (2)$$

Визначимо двосторонньо-обмежений бінарний ряд

$$y(n) = \begin{cases} 1, & s(n) \geq 0 \\ 0, & s(n) < 0 \end{cases}, \quad 0 \leq n \leq N-1$$

і введемо індикаторну функцію F_n для моменту часу n

$$F_n = (y(n) - y(n-1))^2 .$$

Якщо $F_n = 1$, то відбувається перетинання амплітудним значенням сигналу нульового рівня в момент часу n, в протилежному випадку $F_n = 0$. Кількість перетинань амплітудним значенням сигналу нульового рівня для ряду $s(n),\ n = 0,1,\ldots,N-1$ визначається співвідношенням

$$W = \sum_{i=0}^{N-1} F_i .$$

Для процесу, що описується розподілом Гауса з нульовим середнім математичне сподівання кількості перетинань амплітудним значенням сигналу нульового рівня можна представити як

$$\mathrm{E}\{W_1\} = (N-1)\left(\frac{1}{2} - \frac{1}{\pi}\arcsin r(1)\right),$$

де $\mathrm{E}\{\}$ - математичне сподівання, $r(1)$ знаходимо за формулою (2) при $k = 1$.

Тоді

$$r(1) = \cos\left(\frac{\pi \mathrm{E}\{W_1\}}{N-1}\right) = \frac{\int_{-\pi}^{\pi} \cos(\omega) dG(\omega)}{\int_{-\pi}^{\pi} dG(\omega)}. \qquad (3)$$

Для гармонійного сигналу з частотою ω_0 спектральна щільність потужності при $\omega \in [0, \pi]$ повинна задовольняти умові [2]

$$G(\omega) = \frac{A^2}{4}\left[\delta(\omega - \omega_0) + \delta(\omega + \omega_0)\right], \quad (4)$$

де A – амплітуда гармоніки, $\delta(\omega)$ – дельта-функція.

Підставляючи (3) в (4), одержимо рівняння

$$\frac{\pi \mathrm{E}\{W_1\}}{N-1} = \omega_0, \quad (5)$$

тобто частота гармоніки збігається з нормованим значенням математичного сподівання кількості перетинів амплітудним значенням сигналу нульового рівня. Для дискретного сигналу і лінійної частоти, рівняння (5) має вид

$$\frac{\mathrm{E}\{W_1\}}{N-1}\frac{f_d}{2} = f_0,$$

де f_d - частота дискретизації, f_0 - частота гармоніки.

У більш загальному випадку, коли в сигналі присутня деяка домінуюча частота на тлі завад, рівність (5) не виконується, але можна припустити, що нормоване значення кількості перетинів амплітудним значенням сигналу нульового рівня буде прямувати до домінуючої частоти. Якщо не враховувати операцію математичного сподівання, то можна стверджувати, що у випадку коли деяка частота стає домінуючою, величина $W_1 f_d/2(N-1)$ буде мати значення, що відповідає або є близьким до цієї частоти. Таким чином, якщо в спектрі сигналу присутня домінуюча частота, то її можна визначити оцінюванням кількості перетинів амплітудним значенням сигналу нульового рівня.

Даний підхід до виділення домінуючої частоти є ефективним, якщо в сигналі присутня одна гармоніка і відношення сигнал/шум велике. Коли в сигналі присутні кілька гармонік чи амплітуда гармонійної компоненти значно менше шуму, доцільно застосовувати фільтри, які дозволяють виділяти періодичні компоненти сигналу та оцінювати домінуючу частоту в різних частотних смугах.

Для обмеження частотного діапазону, в якому спостерігається частота основного тону, пропонується використовувати полосовий фільтр. Потім до виділеного частотного діапазону почергово застосовуються процедури низькочастної та високочастної фільтрації з використанням нерекурсивних цифрових фільтрів з симетричною імпульсною характеристикою, різницеве рівняння для яких має вигляд

$$y(n) = \sum_{i=0}^{N} h(i)s(n-i),$$

де $h(i)$ - коефіцієнти фільтру, N - порядок фільтру.

Для виділення сигналу в області нижніх частот можна застосувати операцію повторного підсумовування (фільтр нижніх частот), для якого перша сума запишеться у вигляді

$$\Delta s(n) = s(n) + s(n-1),$$

а k-а кінцева сума запишеться як:

$$\Delta^k s(n) = \sum_{j=0}^{k} C_k^j s(n-j) \ .$$

Нехай W_K - кількість перетинів амплітудою сигналу нульового рівня для ряду $\Delta^{k-1}s(n)$. Так як оператор повторного сумування є фільтром нижніх частот, то ряд $\Delta^k s(n)$ буде менш осцилюючим ніж ряд $\Delta^{k-1}s(n)$, отже, і $W_{K+1} < W_K$.

Як фільтр нижніх частот використано кінцево-різницевий оператор ∇, для якого перша різниця має такий вигляд

$$\nabla s(n) = s(n) - s(n-1),$$

а k-а кінцева різниця запишеться як

$$\nabla^k s(n) = \sum_{j=0}^{k} C_k^j (-1)^j s(n-j),\ C_k^j = \frac{k!}{(k-j)!\,j!}.$$

Позначимо W_D - кількість перетинів амплітудою сигналу нульового рівня для ряду $\nabla^{k-1}s(n)$.

Для одержання більш гнучких і точних результатів для оцінювання властивостей сигналу варто одночасно використовувати величини W_D і W_K, послідовно застосовуючи до сигналу процедури фільтрації $\nabla^{k-1}\Delta^{j-1}s(n)$, підраховуючи кількість нульових перетинів.

Отже, для знаходження частоти основного тону за кількістю нульових перетинів можна використати формулу

$$\omega(i) = \frac{f_d}{2M} W(i,j), \qquad (6)$$

$$W_{K_D}(i) = \begin{cases} W_{K_D}(i)+1, & \operatorname{sgn}\left[z_{K_D}(n)\right] \neq \operatorname{sgn}\left[z_{K_D}(n-1)\right] \\ W_{K_D}(i), & \operatorname{sgn}\left[z_{K_D}(n)\right] = \operatorname{sgn}\left[z_{K_D}(n-1)\right] \end{cases},$$

$$W_{K_S}(j) = \begin{cases} W_{K_S}(j)+1, & \operatorname{sgn}\left[z_{K_S}(n)\right] \neq \operatorname{sgn}\left[z_{K_S}(n-1)\right] \\ W_{K_S}(j), & \operatorname{sgn}\left[z_{K_S}(n)\right] = \operatorname{sgn}\left[z_{K_S}(n-1)\right] \end{cases},$$

якщо

$$W(i,j) = \sum_{i,j=1}^{M} W_{K_D}(i) \cdot W_{K_S}(j)$$

де $\omega(i)$ - домінуюча частота, i - порядковий номер K_D в діапазоні $0 \le K_D \le K_{D\max}$, j - порядковий номер K_S у діапазоні $0 \le K_S \le K_{S\max}$, M - число відліків, f_d - частота дискретизації;

$W_{K_S}(j) = \begin{cases} W_{K_S}(j)+1, & \operatorname{sgn}\left[z_{K_S}(n)\right] \neq \operatorname{sgn}\left[z_{K_S}(n-1)\right] \\ W_{K_S}(j), & \operatorname{sgn}\left[z_{K_S}(n)\right] = \operatorname{sgn}\left[z_{K_S}(n-1)\right] \end{cases}$ - кількість нульових перетинів повторно-різнецевої вибірки

$W_{K_D}(i) = \begin{cases} W_{K_D}(i)+1, & \operatorname{sgn}\left[z_{K_D}(n)\right] \neq \operatorname{sgn}\left[z_{K_D}(n-1)\right] \\ W_{K_D}(i), & \operatorname{sgn}\left[z_{K_D}(n)\right] = \operatorname{sgn}\left[z_{K_D}(n-1)\right] \end{cases}$ - кількість нульових перетинів вибірки повторного підсумовування,

$z_{K_D}(n) = \nabla^{K_D}(z_s(n)) = \nabla\left(\nabla^{K_D-1}(z_s)(n)\right),\ n = 1,\ldots,M_D,\ M_D = M_S - K_D,\ K_D = 0,\ldots,K_{\max}$ - сформована повторно-різницева вибірка,

$z_s(n) = \Delta^{K_s} s(n) = \Delta(\Delta^{K_s-1} s(n))$, $n = 1,..., M_s$, $M_s = M - K_s$ - сформована вибірка повторного підсумовування значень $s(n)$.

Метод нормалізації тривалості звучання звукових сигналах на основі частоти основного тону

Для порівняння запису парольної фрази з еталоном слід виконати часову нормалізацію, тобто привести записи парольних фраз до однаковій довжини. Лінійне стиснення або розтягування однієї реалізації слова до величини іншої не вирішує питання внаслідок одної важливої властивості мовного сигналу - нерівномірності його протікання в часі. Ця властивість мовлення виражається у важко контрольованою залежності часу утворення і звучання її елементів від контексту, темпу, діалектних та індивідуальних особливостей диктора. Тому порівняння має спиратися на нелінійну часову нормалізацію.

Нехай дано дві реалізації парольних фраз: $X^{(0)},...,X^{(i)},...,X^{(m)}$ і $Y^{(0)},...,Y^{(j)},...,Y^{(n)}$. Перша реалізація вважається еталонною, друга - новою.

Проведемо нелінійну часову нормалізацію запису мовного сигналу із застосуванням деформуючих функцій (7, 8):

$$\omega_X : \{1,...,l\} \to \{1,...,m\}, \quad (7)$$

$$\omega_Y : \{1,...,l\} \to \{1,...,n\}. \quad (8)$$

При чому:

$$\omega_X(1)=1,\ \omega_Y(1)=1,\ \omega_X(l)=m,\ \omega_Y(l)=n,$$
$$\omega_X(i+1)=\omega_X(i) \text{ або } \omega_X(i)+1,\ \forall\, i=1,...,m-1,$$
$$\omega_Y(j+1)=\omega_Y(j) \text{ або } \omega_Y(j)+1,\ \forall\, j=1,...,n-1. \quad (9)$$

Сегментуюча функція повинна характеризувати сумарну зміну використовуваних нею параметрів мовного сигналу і залежить від двох фреймів: поточного і попереднього. В якості параметрів мовного сигналу ми будемо використовувати частоту основного тону в частотних смугах. Опишемо процедуру знаходження сегментує функції SX (1), ..., SX (i), ..., SX (m) для еталонної реалізації слова.

В кожному фреймі X(i) знаходиться частота основного тону сигналу в частотних смугах: p1 (i), ..., p20 (i); i = 0,1, ...,

Обчислюються модулі кінцевих різниць:

$$\Delta k\ (i) = |\ pki\text{-}pki\text{-}1\ |;\ i = 1, ..., m;\ k = 1, ..., 20. \quad (10)$$

Обчислюються середні різниці:

$$K = 1, ..., 20. \quad (11)$$

Обчислюються середньозважені різниці:

$$\bar{\Delta}_k = \frac{1}{m}\sum_{i=1}^{m} \Delta_k^{(i)};\ i = 1,...,m;\ k = 1,...,n. \quad (12)$$

Контур сегментує функції S_X:

$$\delta_k^{(i)} = \frac{\Delta_k^{(i)}}{\bar{\Delta}_k};\ i = 1,...,m. \quad (13)$$

Аналогічно знаходиться контур сегментуюичих функцій $S_Y(1),...,S_Y(j),...,S_Y(n)$ для нової реалізації слова.

Процедура знаходження деформуючих функцій ω_X, ω_Y реалізується методом динамічного програмування і дає можливість реалізувати внутрішнє нелінійне вирівнювання реалізацій слів в часі.

Спочатку будується матриця відстаней $R = \{ \rho_{i,j} \}$ размірністю $(m \times n)$, далі розраховується матриця $D = \{ d_{i,j} \}$ такої ж розмірності:

$$d_{m,n} = \rho_{m,n};$$
$$d_{i,n} = \rho_{i,n} + d_{i+1,n} , \; i = m-1,...,1;$$
$$d_{m,j} = \rho_{m,j} + d_{m,j+1} , \; j = n-1,...,1;$$
$$d_{i,j} = \rho_{i,j} + \min\{ d_{i+1,j+1} , d_{i+1,j} , d_{i,j+1} \} , \; i = m-1,...,1; \; j = n-1,...,1.$$

Матриця D у свою чергу використовується для знаходження функцій ω_X, ω_Y. Спочатку присвоюються $\omega_X(1)=1$, $\omega_Y(1)=1$. Далі на k-ому кроці знаходять $\omega_X(k+1)$ і $\omega_Y(k+1)$. Можливі чотири випадки:

1. Якщо $\omega_X(k)=m$ і $\omega_Y(k)=n$, то деформуючі функції знайдені;
2. Якщо $\omega_X(k)=m$, а $\omega_Y(k)<n$, то присвоюються: $\omega_X(k+1)=m$, $\omega_Y(k+1)=\omega_Y(k)+1$;
3. Якщо $\omega_X(k)<m$, но $\omega_Y(k)=n$, то присвоюються: $\omega_X(k+1)=\omega_X(k)+1$, $\omega_Y(k+1)=n$;
4. Якщо $\omega_X(k)<m$ и $\omega_Y(k)<n$, то порівнюються $d_{i1,j1}$, $d_{i2,j2}$, $d_{i3,j3}$ для знаходження серед них мінімального відповідних $i_{\min}$, $j_{\min}$.

Тут $i_1=i_2=\omega_X(k)+1$, $i_1=\omega_X(k)$,$j_1=j_3=\omega_Y(k)+1$, $j_2=\omega_Y(k)$.

Потім присвоюємо $\omega_X(k+1)=i_{\min}$, $\omega_Y(k+1)=j_{\min}$.

Знайшовши деформуючі функції ω_X, ω_Y ми можемо для будь-якого відрізка еталонної реалізації мовного сигналу знайти відповідну йому ділянку нової реалізації.

Тестування методу та аналіз результатів

Авторами проведено тестування розробленого методу нормалізації тривалості звучання звукових сигналах на основі частоти основного тону. Демонстраційні результати роботи створеного на основі методу програмного забезпечення наведені на рис. 1-4, де на рис. 1, 3 візуалізовано звукові файли, а на рис. 2, 4 наведено результати часової нормалізації.

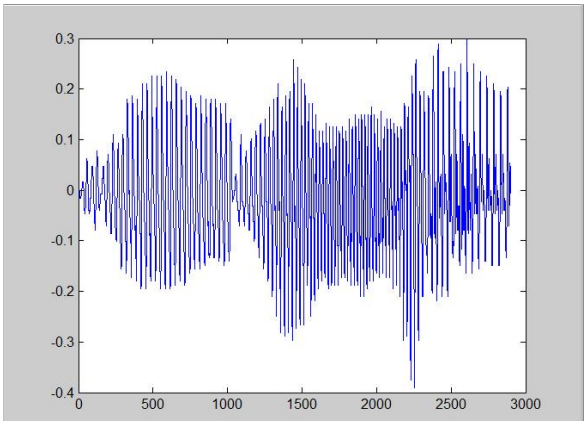

Рисунок 1 – Мовний сигнал до нормалізації

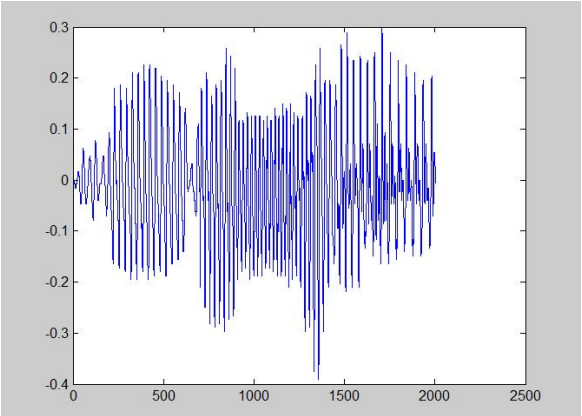

Рисунок 2 – Мовний сигнал після нормалізації (стиснення)

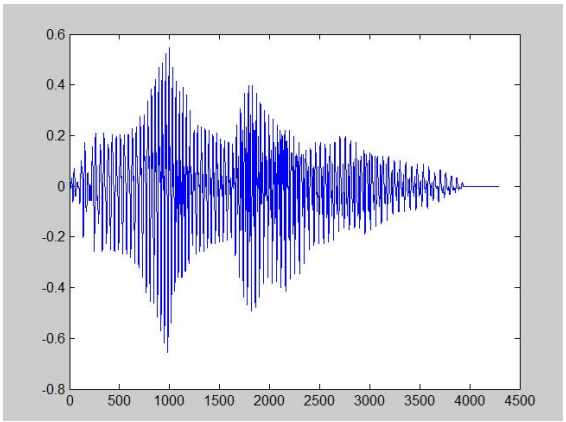

Рисунок 3 - Мовний сигнал до нормалізації

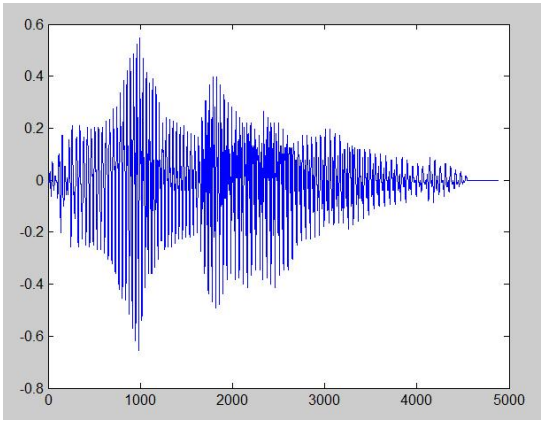

Рисунок 4 – Мовний сигнал після нормалізації (розтягування)

Ефективність створених алгоритмів доведено емпірично, зокрема, проведено експеримент по розпізнаванню 20 осіб за індивідуальними особливостями їх голосів з використанням запропонованого методу часової нормалізації. В результаті сумарна достовірність розпізнавання 20 осіб була вищою 95%, тоді як робота системи без часової нормалізації показала достовірність не вище 70%.

Висновки

Отже, авторами розробленого метод нормалізації тривалості звучання звукових сигналах на основі кореляційного методу виділення частоти

основного тону мовців. Ефективність створених алгоритмів доведено емпірично, зокрема, проведено експеримент по розпізнаванню 20 осіб за індивідуальними особливостями їх голосів з використанням запропонованого методу часової нормалізації. В результаті сумарна достовірність розпізнавання 20 осіб була вищою 95%, тоді як робота системи без часової нормалізації показала достовірність не вище 70%.

Список літератури

1. Айфичер Э.С., Джервис Б.У. Цифровая обработка сигналов: практический подход. Второе издание. Пер. с англ.—М.: Изд. Дом "Вильямс", 2004.—992с.
2. Рамишвили Г.С. Автоматическое опознавание говорящего по голосу. – М.: Радио и связь, 1981. – 224 с.
3. Рабинер Л.Р. Шафер Р.В. Цифровая обработка речевых сигналов: Пер. с англ. – М.: Радио и связь, 1981. – 496 с.
4. Сергиенко А.Б. Цифровая обработка сигналов: Учебное пособие. Второе издание.—СПб.: Питер, 2006.—752 с.