Метод синтезу голосних звуків мовлення за еталонними зразками на основі саундлетів

Євген Є. Федоров

Кафедра автоматизованих систем управління
Донецький національний технічний університет, Донецьк
fee75@mail.ru

Анотація

У тезах викладений метод синтезу голосних звуків мовлення за еталонними зразками на основі саундлетів. Використано материнський і дочірній дискретні й безперервні саундлети та досліджені властивості саундлетних відображень, які дозволяють враховувати структуру квазіперіодичного сигналу та зіставляти зразки голосних звуків мовлення різної довжини. На основі саундлетів і саундлетних відображень розроблений метод створення зразків, метод формування еталонних зразків і модель синтезу голосних звуків за еталонними зразками, які будуть використані в інтелектуальних системах спілкування.

1. Вступ

У сучасних умовах актуальною є розробка інтелектуальних процесорів, призначених для розпізнавання і синтезу мовлення людини та ін., і використовуваних у комп'ютерних системах спілкування. У корені даного завдання лежить проблема побудови ефективних методів, що забезпечують більшу швидкість навчання моделі синтезу, а також більшу адекватність синтезу мовленнєвих сигналів.

Існуючі системи синтезу мовленнєвих образів використовують наступні підходи [1-2]: формантний синтез, синтез на основі коефіцієнтів лінійного прогнозу (КЛП-синтез) і конкатенативний синтез. Формантний синтез і КЛП-синтез опираються на модель мовотворення людини. Модель мовного тракту реалізується у вигляді адаптивного цифрового фільтра. Для формантного синтезу параметри адаптивного цифрового фільтра визначаються формантними частотами [3], а для КЛП-синтеза – КЛП-коефіцієнтами [4]. Кращі результати відносно розбірливості й натуральності звучання мовлення вдається одержати за допомогою конкатенативного синтезу. Конкатенативний синтез здійснюється шляхом склейки потрібних звукових одиниць. У таких системах необхідно застосовувати обробку сигналу для приведення частоти основного тону, енергії й тривалості звукових одиниць до тих, якими повинна характеризуватися синтезоване мовлення. У системах конкатенативного синтезу застосовуються три основних алгоритми: TD-PSOLA (здійснюється масштабування звукової одиниці за часом), FD-PSOLA (здійснюється масштабування звукової одиниці за частотою), LP-PSOLA (здійснюється масштабування сигналу помилки прогнозу за часом з наступним застосуванням фільтра із КЛП-коефіцієнтами). Недоліком конкатенативного синтезу є необхідність зберігання великої кількості звукових одиниць, у зв'язку з чим виникає завдання їх більш ощадливого подання [5].

Метою роботи є розробка методу синтезу голосних звуків мовлення, що базується на саундлетах і формованих на їхній основі саундлетних відображеннях.

Для досягнення поставленої мети необхідно:

1. Розробити метод створення сімейства зразків голосних звуків.
2. Сформувати сімейства материнських і дочірніх саундлетів, що характеризують зразки голосних звуків.
3. Формалізувати саундлетні відображення, що діють між сімействами зразків і саундлетів, результатом яких є зразок, що перебуває в заданому амплітудно-тимчасовому вікні.
4. Розробити метод формування еталонних зразків на основі сімейства дискретних саундлетів і саундлетних відображень.
5. Розробити модель синтезу голосних звуків за еталонними зразками на основі сімейства саундлетів і саундлетних відображень.

2. Метод створення сімейства зразків голосних звуків

Зразком голосного звуку мовлення назвемо ділянку голосного звуку в мовленнєвому сигналі, що розташований між сусідніми піковими значеннями й має довжину відповідну квазіперіоду.

При формуванні зразка експертом уводяться ліва й права границі N^l, N^r голосного звуку в сигналі f.

Після задання границь N^l, N^r на множині $\{N^l,...,N^r\}$ сигналу f обчислюється функція автокореляції, за допомогою якої визначається довжина періоду основного тону N^{FT} голосного звуку.

Для формування зразка як структуроутворюючого елемента голосного звуку на множині $\{N^l,...,N^r\}$ сигналу f розбивається на ділянки на основі обчисленої довжини періоду основного тону N^{FT} відповідно до наступного правила

$$N_0^{\max} = \arg\max_n f(n), \qquad (1)$$

$$n \in \{N^l - 0.5 \cdot N^{FT},..., N^l + 0.5 \cdot N^{FT}\},$$

$$N_{i-1}^{\max} \le N^r \Rightarrow \left(N_i^{\min} = N_{i-1}^{\max}\right) \wedge$$

$$\wedge\left(N_i^{\max} = \arg\max_n f(n) \right), \quad (2)$$

$$n \in [N_i^{\min} + 0.5 \cdot N^{FT}, N_i^{\min} + 1.5 \cdot N^{FT}],$$

На основі цієї розбивки формується кінцеве сімейство зразків, які описуються множиною цілочисельних обмежених фінітних дискретних функцій $X = \{x_i \mid i \in \{1,...,I\}\}$, у вигляді

$$x_i(n) = \begin{cases} f(n), & n \in \{N_i^{\min},...,N_i^{\max}\} \\ 0, & n \notin \{N_i^{\min},...,N_i^{\max}\} \end{cases}, \quad (3)$$

$$A_i^{\min} = \min_n f(n), \quad (4)$$

$$A_i^{\max} = \max_n f(n), \quad (5)$$

$$n \in \{N_i^{\min},...,N_i^{\max}\},\ i \in \{1,...,I\},,$$

Для подальшого зіставлення зразків між собою при формуванні еталонних зразків необхідно привести їх до подібності (тобто до єдиного прямокутного амплітудно-тимчасового вікна, у яке точно вписана тільки та частина зразка, що перебуває на компактному носії). Для цього в статті розробляються материнський і дочірній саундлети.

3. Створення сімейства материнських дискретних саундлетів

Материнським саундлетом зразка голосного звуку мовлення назвемо зразок, що перемістить за часом й амплітудою в лівий нижній кут позитивної площини.

Материнський дискретний саундлет зразка голосного звуку мовлення представлений у вигляді цілочисельної обмеженої фінітної дискретної функції

$$s^m(n) = \begin{cases} x(n+b0) - d0, & n \in \{0,...,N\} \\ 0, & n \notin \{0,...,N\} \end{cases}, \quad (6)$$

$$b0 = N^{\min},\ d0 = A^{\min},$$

$$N = N^{\max} - N^{\min},\ A = A^{\max} - A^{\min},$$

де $F0$ – перетворення, що переводить зразок у материнський саундлет, $b0$, $d0$ – параметри зрушення функції x за часом й амплітудою, $A^{\min}, A^{\max}$ – мінімальне й максимальне значення функції x на компактному носії $\{N^{\min},...,N^{\max}\}$.

Таким чином, частина материнського саундлета, що перебуває на компактному носії $\{0,...,N\}$, точно вписана в амплітудно-тимчасове вікно висотою A й шириною N.

Визначимо кінцеве сімейство материнських дискретних саундлетів зразків голосного звуку мовлення як $S^m = \{s^m\}$, причому всі функції S^m обмежені знизу й зверху числами 0 і A відповідно.

Від материнського дискретного саундлета породимо материнський безперервний саундлет.

4. Створення сімейства материнських безперервних саундлетів

Материнський безперервний саундлет ψ^m отриманий з материнського дискретного саундлета s^m на основі лінійної інтерполяції.

Материнський безперервний саундлет зразка голосного звуку мовлення представлений у вигляді речовиннозначної обмеженої фінітної безперервної функції

$$\psi^m(t) = \begin{cases} \sum_{n=-1}^{N} \chi_{(t_n,t_{n+1})}(t)\hat{s}^m(n) + \\ + \sum_{n=-1}^{N+1} \chi_{\{t_n\}}(t)s^m(n), & t \in [-\Delta t, T+\Delta t] \\ 0, & t \notin [-\Delta t, T+\Delta t] \end{cases}, \quad (6)$$

$$\hat{s}^m(n) = s^m(n) + \frac{s^m(n+1) - s^m(n)}{\Delta t}(t - t_n),$$

$$T = N\Delta t,\ t_n = n\Delta t,\ \chi_B(t) = \begin{cases} 1, & t \in B \\ 0, & t \notin B \end{cases},$$

де Δt – крок квантування за часом.

Таким чином, частина материнського саундлета, що перебуває на компактному носії $[-\Delta t, T+\Delta t]$, точно вписана в амплітудно-тимчасове вікно висотою A й шириною $T + 2\Delta t$.

Визначимо кінцеве сімейство материнських безперервних саундлетів зразків голосного звуку мовлення як $\Psi^m = \{\psi^m\}$, причому всі функції Ψ^m обмежені знизу й зверху числами 0 і A відповідно.

Від материнського безперервного саундлета породимо дочірній безперервний саундлет, що описує зразок голосного звуку мовлення, що перебуває в заданому амплітудно-тимчасовому вікні.

5. Створення сімейства дочірніх безперервних саундлетів

Дочірнім саундлетом назвемо зрушений і масштабований за часом й амплітудою материнський саундлет.

Дочірній безперервний саундлет представлений у вигляді речовиннозначної обмеженої фінітної безперервної функції

$$\psi^c(t) = \begin{cases} 0, & t \le \breve{T}^{\min} - \Delta t \\ \left(d + c\psi^m\left(\frac{\breve{T}^{\min} - b}{a} \right) \right) & \\ \left(\frac{t - (\breve{T}^{\min} - \Delta t)}{\Delta t} \right), & t \in [\breve{T}^{\min} - \Delta t, \breve{T}^{\min}] \\ d + c\psi^m\left(\frac{t-b}{a} \right), & t \in [\breve{T}^{\min}, \breve{T}^{\max}] \\ \left(d + c\psi^m\left(\frac{\breve{T}^{\max} - b}{a} \right) \right) & \\ \left(\frac{(\breve{T}^{\max} + \Delta t) - t}{\Delta t} \right), & t \in [\breve{T}^{\max}, \breve{T}^{\max} + \Delta t] \\ 0, & t \ge \breve{T}^{\max} + \Delta t \end{cases}, \quad (8)$$

$$a = \frac{\breve{T}^{\max} - \breve{T}^{\min}}{T},\ b = \breve{T}^{\min},$$

$$c = \frac{\breve{A}^{\max} - \breve{A}^{\min}}{\tilde{A}^{\max} - \tilde{A}^{\min}},\ d = \breve{A}^{\min}$$

$$\tilde{A}^{\max} = \max_t \psi^m\left(\frac{t-b}{a}\right),\ \tilde{A}^{\min} = \min_t \psi^m\left(\frac{t-b}{a}\right),$$

$$t \in [\breve{T}^{\min}, \breve{T}^{\max}],$$

де a, c – параметри масштабування функції ψ^m за часом й амплітудою, b, d – параметри зрушення функції ψ^m за часом й амплітудою, $\breve{A}^{\min}, \breve{A}^{\max}$ – задане мінімальне й максимальне значення функції ψ^c на компактному носії $[\breve{T}^{\min}, \breve{T}^{\max}]$.

Таким чином, частина дочірнього саундлета, що перебуває на компактному носії $[\breve{T}^{\min} - \Delta t, \breve{T}^{\max} + \Delta t]$, точно вписана в задане амплітудно-тимчасове вікно висотою $\breve{A}^{\max} - \breve{A}^{\min}$ й шириною $\breve{T}^{\max} - \breve{T}^{\min} + 2\Delta t$.

Визначимо кінцеве сімейство дочірніх безперервних саундлетов зразків голосного звуку мовлення як $\Psi^c = \{\psi^c\}$, причому всі функції Ψ^c мають однаковий компактний носій $[\breve{T}^{\min} - \Delta t, \breve{T}^{\max} + \Delta t]$ й однакові мінімальні й максимальні значення $\breve{A}^{\min}$, $\breve{A}^{\max}$ на ньому.

Від дочірнього безперервного саундлета породимо дочірній дискретний саундлет.

6. Створення сімейства дочірніх дискретних саундлетів

Дочірній дискретний саундлет s^c отриманий з дочірнього безперервного саундлета ψ^c шляхом дискретизації.

Дочірній дискретний саундлет представлений у вигляді цілочисельної обмеженої фінітної дискретної функції

$$s^c(n) = round(\psi^c(n\Delta t)), \quad (9)$$

$$n \in \{\breve{N}^{\min}, ..., \breve{N}^{\max}\},$$

$$\breve{N}^{\min} = \breve{T}^{\min}/\Delta t, \breve{N}^{\max} = \breve{T}^{\max}/\Delta t,$$

де *round* – функція, що округляє число до найближчого цілого.

Таким чином, частина дочірнього саундлета, що перебуває на компактному носії $\{\breve{N}^{\min}, ..., \breve{N}^{\max}\}$, точно вписана в задане амплітудно-тимчасове вікно висотою $\breve{A}^{\max} - \breve{A}^{\min}$ й шириною $\breve{N}^{\max} - \breve{N}^{\min}$.

Визначимо кінцеве сімейство дочірніх дискретних саундлетов зразків голосного звуку мовлення як $S^c = \{s^c\}$, причому всі функції S^c мають однаковий компактний носій $\{\breve{N}^{\min}, ..., \breve{N}^{\max}\}$ й однакові мінімальні й максимальні значення $\breve{A}^{\min}$, $\breve{A}^{\max}$ на ньому.

Для перетворення зразка з метою приведення його до подібності (однакового амплітудно-тимчасового вікна) формалізуємо відображення між зразками, материнськими саундлетами й дочірніми саундлетами.

7. Формалізація саундлетних відображень

Саундлетним відображенням назвемо перетворення, що переводить зразок у материнський дискретний саундлет, материнський дискретний саундлет у материнський безперервний саундлет, материнський безперервний саундлет у дочірній безперервний саундлет, дочірній безперервний саундлет у дочірній дискретний саундлет шляхом лінійної інтерполяції, зсув й масштабування за часом й амплітудою, дискретизації.

Перетворення $F0$, що здійснює зсув функції x, що описує зразок, за часом й амплітудою в лівий нижній кут позитивної площини для одержання материнського дискретного саундлета s^m, представимо у вигляді саундлетного відображення $F0: X \to S^m$.

Перетворення $F1$, що створює з материнського дискретного саундлета s^m материнський безперервний саундлет ψ^m, представимо у вигляді саундлетного відображення $F1: S^m \to \Psi^m$.

Перетворення $G1$, що здійснює зсув й масштабування материнського безперервного саундлета ψ^m за часом й амплітудою для одержання дочірнього безперервного саундлета ψ^c, представимо у вигляді саундлетного відображення $G1: \Psi^m \to \Psi^c$.

Перетворення $F2$, що створює шляхом дискретизації з дочірнього безперервного саундлета ψ^c дочірній дискретний саундлет s^c, представимо у вигляді саундлетного відображення $F2: \Psi^c \to S^c$.

Композиція перетворень $F0$, $F1$, $G1$, $F2$ представлена у вигляді $F = F2G1F1F0$.

Таким чином, перетворення, що здійснює перехід від функції x, що описує зразок, до дочірнього дискретного саундлету s^c, представимо у вигляді саундлетного відображення $F: X \to S^c$, яке має наступні властивості

1. Однаковий компактний носій у всіх дочірніх саундлетів

$$\forall \hat{x} \in X \quad \forall \breve{x} \in X \quad \text{supp}F\hat{x} = \text{supp}F\breve{x}, \quad (10)$$

2. Однакові мінімальні та максимальні значення на компактному носії у всіх дочірніх саундлетів

$$\forall \hat{x} \in X \quad \forall \breve{x} \in X \quad \left(\min_{n \in \text{supp}F\hat{x}} (F\hat{x})(n) = \min_{n \in \text{supp}F\breve{x}} (F\breve{x})(n)\right) \wedge$$

$$\wedge \left(\max_{n \in \text{supp}F\hat{x}} (F\hat{x})(n) = \max_{n \in \text{supp}F\breve{x}} (F\breve{x})(n)\right), \quad (11)$$

Обмеження 1-2 забезпечують єдине прямокутне амплітудно-тимчасове вікно для всіх отриманих дочірніх саундлетів, у яке точно вписана тільки та частина цих саундлетів, що перебуває на компактному носії.

На основі введених сімейств саундлетів і саундлетних відображень сформуємо еталонні зразки голосних звуків мовлення.

8. Метод формування еталонних зразків

Нехай дана кінцева сукупність навчальних зразків голосного звуку, що описується множиною цілочисельних обмежених фінітних дискретних функцій $X = \{x_i \mid i \in \{1,...,I\}\}$, причому $A_i^{\min}, A_i^{\max}$ – мінімальне й максимальне значення функції x_i на компактному носії $\{N_i^{\min},...,N_i^{\max}\}$.

Для зіставлення елементів множини X між собою для кожної функції x_i, що описує навчальний зразок, формується відповідна йому кінцева множина дочірніх дискретних саундлетів S^c, що перебувають у тому самому амплітудно-тимчасовому вікні, що й ця функція у вигляді

$$\forall x_i \in X \;\; \exists S^c = \{s_r^c \mid r \in \{1,...,I\}\} \;:\; s_r^c = Fx_r, \qquad (12)$$

Обчислюється нормована відстань між функцією, що описує навчальний зразок, і дочірнім дискретним саундлетом у вигляді

$$d_{ir} = \frac{\rho_p(x_i, s_r^c)}{(A_i^{\max} - A_i^{\min})\sqrt[p]{(N_i^{\max} - N_i^{\min} + 1)}}, \qquad (13)$$

$$\rho_p(x_i, s_r^c) = \sqrt[p]{\sum_{m \in \mathbb{Z}} | x_i(m) - s_r^c(m) |^p}, \; i, r \in \{1,...,I\}$$

Далі здійснюється вибір множині функцій H, що описують еталонні зразки, з множині функцій X, які описують навчальні зразки, на основі матриці нормованих відстаней $[d_{ir}]$.

9. Процедура вибору підмножини еталонних зразків з множини навчальних зразків

Приведемо етапи процедури вибору підмножини еталонних зразків з множини навчальних зразків на основі матриці $[d_{ir}]$

1. Створити точково кінцеве покриття C множини номерів навчальних зразків $B0 = \{1,...,I\}$ у вигляді

1.1. $C = \{C_i\}$, $C_i = \{r \mid d_{ir} < \varepsilon, r \in B0\}$, $i \in B0$, $0 < \varepsilon \le 1$.

1.2. $\forall i \in B0 \;\; |C_i| < \delta \Rightarrow C_i = \{i\}$, $1 < \delta < I$,

причому ε, δ задаються експертом.

2. Створити множину $B1$ із номерів елементів покриття у вигляді $B1 = \{i \parallel C_i | > 1\}$.

3. Створити множину $B2$ із елементів $B1$ у вигляді $B2 = \left\{ i \middle| (i \in B1) \wedge \left(\underset{m \in B1, i \ne m}{\wedge} (C_i \ne C_m \wedge C_i \not\subset C_m) \right) \vee \left(C_i = C_m \wedge \sum_{z \in C_i} d_{iz} > \sum_{z \in C_m} d_{mz} \right) \right\}$.

4. Створити множину $B3$ із елементів $B2$

4.1. $i = 2$, $E1 = \varnothing$.

4.2. Якщо $j = b2_i \wedge C_j \subset \underset{m \in V}{\cup} C_m$, те $E1 = E1 \cup \{j\}$, де $V = (B2 \cap E1) \setminus \{j\}$.

4.3. Якщо $i < |B2|$, те $i = i + 1$, перехід до кроку 4.2, інакше $B3 = B2 \setminus E1$.

5. Створити кінцеву множину еталонних зразків H у вигляді $H = \{h_k \mid h_k = x_{i_k}, i_k \in B4\}$, $B4 = B3 \cup \left(B0 \setminus \underset{m \in B3}{\cup} C_m \right)$

6. Створити підпокриття $\tilde{C}$ у вигляді $\tilde{C} = \{\tilde{C}_k \mid \tilde{C}_k = C_{i_k}, i_k \in B4\}$, $\tilde{C} \subset C$.

На основі введених сімейств саундлетів і саундлетних відображень і сформованої множини еталонних зразків і підпокриття $\tilde{C}$ створимо модель синтезу голосного звуку за еталонними зразками.

10. Модель синтезу голосного звуку за еталонними зразками

Модель синтезу голосного звуку за еталонними зразками створюється на основі детермінованого кінцевого автомата. Детермінований кінцевий автомат, що синтезує деякий голосний звук з обліком його квазіперіодичної структури, множини еталонних зразків H і підпокриття $\tilde{C}$, представлений у вигляді графа на рис. 1.

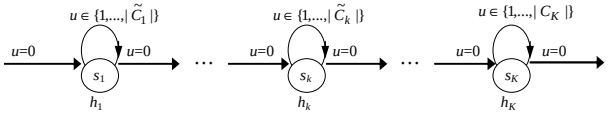

Рисунок 1. Детермінований кінцевий автомат, що синтезує голосний звук

11. Висновки

У роботі вдосконалений підхід до синтезу голосних звуків, що відрізняється тим, що дозволяє узагальнювати зразки одного звуку різної довжини й з різним розмахом амплітуд, що підвищує ефективність синтезу голосних звуків мовлення. Одержав подальший розвиток метод створення множині еталонних зразків, що відрізняється тим, що засновано на сімействах саундлетів і саундлетних відображень, що підвищує ефективність процедури формування еталонних зразків. На основі сімейств саундлетів і саундлетних відображень удосконалена модель синтезу голосних звуків, що відрізняється тим, що дозволяє зіставляти зразки різної довжини, що підвищує ефективність синтезу корисних звуків.

Розроблено метод побудови моделі синтезу голосних звуків за еталонними зразками на основі сімейств саундлетів і саундлетних відображень, що дозволяє скоротити кількість еталонних зразків. Створені алгоритми можуть бути використані для рішення завдань, пов'язаних з конкатенативним синтезом мовлення.

12. Література

[1] Бондарев В.Н., Аде Ф.Г., *Искусственный интеллект*, Изд-во СевНТУ, Севастополь, 2002, 615 с.

[2] Потапова Р.К., *Речь: коммуникация, информация, кибернетика*, Радио и Связь, Москва, 1997, 528 с.

[3] *Искусственный интеллект [в 3-х кн.]. – Кн.. 1. Системы общения и экспертные системы: Справочник* / Под ред. Э.В. Попова, Радио и связь, Москва, 1990, 464 с.

[4] Rabiner L.R., Jang B.H., *Fundamentals of speech recognition*, Prentice Hall PTR, Englewood Cliffs, 1993, 507 p.

[5] Винцюк Т.К., *Анализ, распознавание и интерпретация речевых сигналов*, Наукова думка, Киев, 1987, 264 с.