Автоматизована система отримання стенограм засідань

Ю.Г.Кривонос, Ю.В.Крак, О.В.Бармак, О.С.Загваздін

Інститут кібернетики ім. В.М.Глушкова НАНУ
03680, МСП, Київ, проспект академіка Глушкова, 40
krak@unicyb.kiev.ua, barmak@svitonline.com.ua, alex.zagvazdin@gmail.com

Abstract

A new approach for stenography process is suggested. Especial attention is paid to the usability of the user interface and sound signal preprocessing. Then a comparison of practical efficiencies of the technology prototype and existing approaches is made.

1. Вступ

Оскільки стенографування засідань є задачею, яка може бути досить зручно розподілена між багатьма виконавцями, система автоматизованого стенографування має підтримувати як однокористувацький так і багатокористувацький режими роботи. На даний момент в світі існує декілька систем розподіленого стенографування, проте їм властиві декілька недоліків, яких позбавлена запропонована система. Основним недоліком існуючих систем стенографування засідань є те що вони пред'являють високі вимоги до апаратного забезпечення, на якому працює серверна частина система і суттєвих затрат на впровадження таких систем. Такі системи також вимагають постійного адміністрування. Очевидно, що ці обмеження роблять майже неможливими впровадження систем стенографування у невеликих організаціях, організаціях з обмеженим IT бюджетом, та для індивідуальних користувачів. Наявність серверної частини в таких системах також майже унеможливлює мобільну роботу з системою і вимагає постійного підключення до локальної мережі. Наявність текстової стенограми засідань є важливою частиною роботи для багатьох установ. Зазвичай процес створення і розшифровки стенограм є досить тривалим і пришвидшувати його за допомогою нарощення персоналу вважається неефективним. Для автоматизації процесу отримання стенограм засідань пропонується система розподіленого комп'ютерного стенографування.

Запропонована система призначена автоматизувати і спростити роботу індивідуальних операторів стенографістів і груп стенографістів в організаціях різного рівня і є продовженням роботи над розподіленою системою автоматизованого стенографування [1,2]. Очевидним є той факт, що навички роботи з комп'ютерами операторів стенографістів в більшості організацій є обмеженими, що накладає особливі вимоги на інтерфейс користувача і ергономіку системи автоматизованого стенографування. Зокрема, якомога більша кількість стандартних операцій в системі має бути автоматизована і не вимагати дій від користувача, а набір операцій, в яких приймає участь користувач має бути обмеженим і складатися з невеликого числа чітко зрозумілих операцій. При цьому система повинна залишатися гнучкою і підтримувати налаштування для того, щоб зробити роботу більш зручною.

2. Вимоги до системи стенографування

Ефективна система автоматизованого створення стенограм засідань має задовольняти наступним вимогам:

- отримувати і зберігати мовний голосовий сигнал, підтримувати більшість існуючих форматів зберігання звукової інформації;
- реалізовувати попередню цифрову обробку звукового сигналу, зокрема позбавлення його від сторонніх шумів;
- автоматично розбивати сигнал на сегменти, довжина яких була б зручною для обробки стенографістом;
- підтримувати однокористувацький і багатокористувацький режими роботи системи, при цьому система не має вимагати комплексного впровадження і адміністрування в рамках організації;
- в багатокористувацькому режимі ефективно розподіляти сегменти між операторами-стенографістами.

Очевидним є факт, що рівень комп'ютерної грамотності операторів-стенографістів є не дуже високим, що накладає особливі вимоги на ергономіку системи і інтерфейс користувача.

3. Характеристики запропонованої системи

Запропонована система має наступні функціональні властивості:

- Інтерфейс користувача відповідає основним вимогам, які висуваються до ергономіки і зручності інтерфейсу програмного продукту. Зокрема, згідно з класичним дослідженням Джорджа Міллера про короткочасну пам'ять людини, яке стверджує, що людина може концентрувати увагу одночасно лише на 7+/-2 об'єктах, кількість основних команд, які доступні користувачу при роботі з голосовим сигналом зведена до 8, а оптимальна довжина сегмента визначена на рівні 5-9 слів. До того ж для всіх основних команд в інтерфейсі реалізовано «гарячі клавіші», щоб користувач міг керувати системою не відриваючи рук від клавіатури при набиранні тексту.
- Система виконує ефективне розбиття звукового сигналу на еквівалентні сегменти. Таке розбиття реалізовано за наступним принципом: по сигналу проходять вікном визначеної довжини і визначають

частини сигналу, які відповідають паузам у мовленнєвому потоці (середньоквадратичне відхилення в таких сегментах не перевищує деякої заданої порогової величини), за цими паузами визначаються границі сегментів, при цьому довжина сегмента не є меншою за деяку визначену величину. Довгі паузи вирізаються з сигналу.

- Система працює як в однокористувацькому, так і в багатокористувацькому режимах.
- Серед комп'ютерів, на яких встановлена система і які об'єднані в мережу один визначається як головний; він виконує роль сервера, забезпечуючи отримання сигналу, його обробку і розбиття на сегменти і розподілення сегментів між рештою комп'ютерів – таким чином система майже не вимагає впровадження і адміністрування.
- Система виконує попередню цифрову обробку звукового сигналу: фільтрує його від сторонніх шумів за допомогою, спеціальним чином підібраних, вейвлет перетворень, здатна змінювати амплітуду сигналу, частоту основного тону сигналу та швидкість його відтворення, здатна створювати ефекти «об'ємного» звучання.
- В багатокористувацькому режимі система ефективно розподіляє сегменти між операторами-стенографістами. На головному комп'ютері системи сегменти організовано у вигляді черги, при звільнені чергового оператора-стенографіста йому надсилається наступний сегмент.

Інтерфейс АРМ оператора-стенографіста представлено на наступному малюнку:

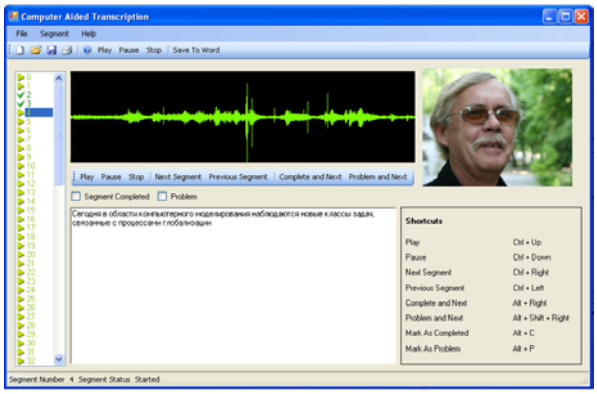

Рис. 1. Інтерфейс користувача системи

4. Попередня обробка звукового сигналу

В системі для попередньої обробки сигналу, який містить шум використовується методика, яка разом з адаптацією до шуму враховує акустичні особливості широких фонетичних класів і складається з п'яти етапів:

- обчислення порогів за зразком шуму (навчання шуму);
- відмічання фреймів сигналу;
- визначення границь слів;
- видалення шуму з вейвлет-образу сигналу;
- відтворення сигналу за оновленими коефіцієнтами.

На етапі навчання шуму виконується вейвлет розклад сигналу $s(n)$, що містить зразок шуму, по рівням $j = \overline{j_{\min}; j_{\max}}$, потім цей сигнал розбивається на фрейми. Для кожного *s*-го фрейму обраховуються енергії по всім рівням розкладу *m:*

$$E_s(m) = \sum_{n=(s-1)\Delta N}^{s\Delta N} d_{mn}^2, \quad m \in \overline{j_{\min}; j_{\max}}$$

На основі отриманих енергій будується модифікована міра контрастності:

$$C(m,s) = \lg\left(\frac{E_s(m)}{\sum_{j=j_{\min}}^{m} E_s(j)}\right), \quad m \in \overline{j_{\min}+1; j_{\max}}$$

По всім фреймам сигналу на кожному *m*-му рівні розкладу отримуються пороги:

$$\alpha(m) = \min_{s \in F_{noise}} C(m,s), \quad \beta(m) = \max_{s \in F_{noise}} C(m,s),$$

Тут F_{noise} - множина номерів фреймів, на які розбивається сигнал, що містить лише шум і використовується для навчання. Окрім порогів обраховуються усереднені енергетичні характеристики шуму:

$$AverE(m) = \frac{\sum_{s \in M_{noise}} E_s(m)}{|F_{noise}|\Delta N}$$

Отримані пороги і усереднені енергії заносяться в базу даних.

На етапі позначення фреймів кожен фрейм відноситься до одного з наступних класів:

- лише шум *(Noise)*;
- вокалізований звук *(Voc)*;
- шумний глухий щілинний або смично-щілинний звук *(Sh)*;
- шумний глухий смичний звук *(P)*.

Для проведення класифікації виділяються дві множини масштабів:

M_{voc} - відповідає смузі частот основного тону (100-300 Гц), де зростають значення функції *C(m,s)* для фреймів, що містять локалізований звук.

M_{sh} - відповідає високочастотній частині спектру (більше 4 кГц), в якій зосереджена енергія шумних звуків.

Класифікація фреймів відбувається за наступними правилами:

$$\forall m:\ \alpha(m) \le C(m,s) \le \beta(m) \to s \in Noise \vee P$$

$$(\forall m \in M_{sh} : \alpha(m) > C(m,s)) \wedge (\exists n \in M_{voc} : C(n,s) > \beta(n)) \to$$
$$\to s \in Voc$$

$$\exists m \in M_{sh} : \beta(m) < C(m,s) \to s \in Sh$$

де $M_{voc} = \{m : m_{voc} \le m \le j_{\max}\}$ $M_{sh} = \{m : j_{\min} \le m \le m_{sh}\}$,

На основі класифікації фреймів будується функція їх позначення:

$$Mark(s) = \begin{cases} 0, & s \in Noise \vee P \\ 1, & s \in Voc \\ 2, & s \in Sh \end{cases}$$

Наступний етап – визначення границь слова. Номери відліків, які є лівою і правою границями слова визначаються згідно правил:

$$\exists N_l : (\forall s < N_l \ \ Mark(s) = 0) \wedge (Mark(N_l) \neq 0) \to L = N_l \Delta N$$
$$\exists N_r : (\forall s : N_r < s \le N_r + L_{\max} Mark(s) = 0) \wedge (Mark(N_r) \neq 0) \to R = N_r \Delta N$$

На етапі видалення шумової складової з вейвлет образу сигналу проводиться оновлення вейвлет коефіцієнтів. Значення d_{im} обраховуються для кожного фрейма $(s-1)\Delta N \le m < s\Delta N$ з урахуванням його позначення

$$Mark(s) = 0 \vee Mark(s) = 3 \to \forall i : j_{\min} \le i \le j_{\max}$$

$$Mark(s) = 1 \to$$

$$\tilde{d}_{im} = \begin{cases} d_{im} - \sqrt{AverE(m)}, & (d_{im}^2 > AverE(m)) \wedge (m_{sh} < i \le j_{\max}) \\ 0, & (d_{im}^2 \le AverE(m)) \vee (i \le m_{sh}) \end{cases}$$

$$Mark(s) = 2 \to$$

$$\tilde{d}_{im} = \begin{cases} d_{im} - \sqrt{AverE(m)}, & (d_{im}^2 > AverE(m)) \wedge (j_{\min} \le i \le m_{sh}) \\ 0, & (d_{im}^2 \le AverE(m)) \vee (i \ge m_{voc}) \end{cases}$$

Після цього виконується останній етап – відтворення сигналу по оновлених вейвлет коефіцієнтах.

Зміна частоти основного тону і тривалості сигналу для відтворення ефектів повільнішого і швидшого мовлення досягається за допомогою PSOLA-подібних алгоритмів [12]. Для реалізації цих алгоритмів спочатку розв'язується задача знаходження границь періодів псевдоперіодичності у мовному сигналі (пітч-періодів). Для цього вхідний звуковий сингал пропускається через низькочастотний і високочастотний фільтри зі скінченими імпульсними характеристиками. Результат застосування фільтрів до складу «ма» наведено на рис. 3.

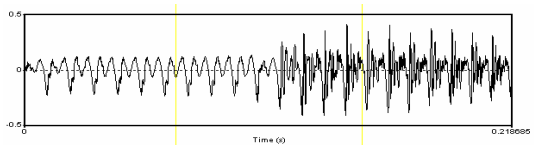

Рис 2. Склад «ма» до застосування фільтрів.

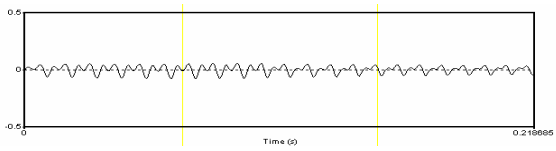

Рис. 3. Склад «ма» після фільтрації.

Далі, для покращення гладкості сигналу кожен елемент вектору вхідного сигналу замінюється на зважене середнє чотирьох оточуючих елементів за формулою: (2.24)

$$d[i] = \frac{3x[i-2] + x[i-1] - x[i+1] - 3x[i+2]}{10}$$

і до отриманого сигналу застосовується медіанне згладження порядку $n = 199$ (кожен елемент вектора замінюємо на медіану вектора, що складається з n елементів, що оточують поточний елемент). Вигляд сигналу після згладження зображено на рис. 4.

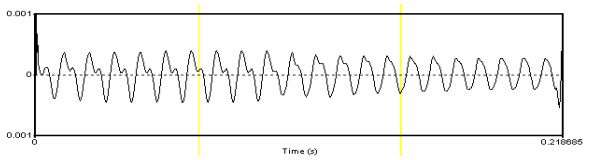

Рис. 4. Склад «ма» після фільтрації і медіанного згладження.

Після цього в отриманому сигналі знаходяться точки, де послідовність, що складається з елементів вектору сигналу змінює знак з – на + і такі точки позначаються як границі пітч-періодів. Серед визначених границь знаходяться і вилучаються точки, що знаходяться занадто близько одна від одної, а для ділянок сигналу, де немає псевдоперіодичності призначаються умовні границі з деяким сталим інтервалом.

Вхідний сигнал можна представити у вигляді функції періодів основного тону $x_i[n]$:

$$x[n] = \sum_{i=-\infty}^{\infty} x_i[n - t_a[i]]$$

де $t_a[i]$ - границі періодів псевдоперіодичності сигналу, тобто різниця між двома сусідніми границями $P_a[i] = t_a[i] - t_a[i-1]$ дорівнює періоду основного тону в момент часу $t_a[i]$. Пітч-період визначимо як вхідний сигнал помножений на віконну функцію

$$x_i[n] = w_i[n]x[n]$$

де вікна w_i задовольняють умові

$$\sum_{i=-\infty}^{\infty} w_i[n - t_a[i]] = 1$$

що досягається використанням віконних функцій типу Хеннінга або трапецієвидним вікном довжиною 2 періоди основного тону.

В результаті роботи алгоритму маємо отримати сигнал $y[n]$, який має однакові з $x[n]$ спектральні характеристики, але відрізняється від нього за основним тоном і/або тривалістю. Щоб досягти цього, замінюємо аналітичні границі пітч-періодів $t_a[i]$ границями $t_b[i]$, а аналітичні періоди основного тону $x_i[n]$ періодами $y_i[n]$, де

$$y[n] = \sum_{j=-\infty}^{\infty} y_j[n - t_b[j]]$$

Таким чином тепер лише достатньо задати границі $t_b[i]$, які відповідають тривалості і основному тону, які маємо отримати. Результуючий період основного тону $y_i[n]$ отримуємо підстановкою найближчого відповідного аналітичного періоду $x_i[n]$.

Графічно роботу алгоритму представлено на рис. 5.

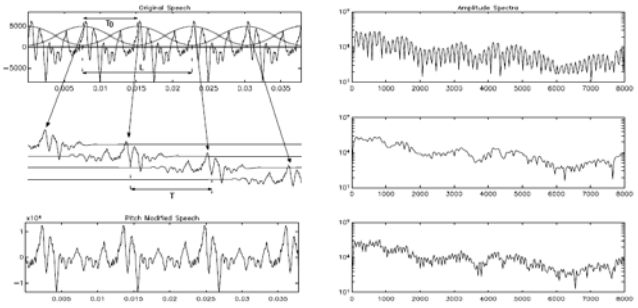

Рис. 5. Схема роботи алгоритму для модифікації тривалості і основного тону.

Зміна амплітуди звукового сигналу для голоснішого чи тихішого відтворення запису досягається множенням елементів вектору сигналу на відповідні коефіцієнти підсилення.

5. Результати експерименту

Після реалізації прототипу системи було проведено експеримент, метою якого було порівняти ефективність роботи одного стенографіста і групи стенографістів при використанні системи і при використанні лише традиційних засобів (таких як Windows Media Player для відтворення звукового сигналу і Microsoft Office Word для набирання тексту стенограми). В якості вхідного звукового файлу для експерименту було обрано запис засідання вченої ради Інституту Кібернетики НАН України з приводу захисту докторської дисертації тривалістю 2 години. Отримано наступні результати:

- При роботі одного стенографіста на стенографування запису з використанням систему було витрачено близько 6 годин. Для стенографування такого запису при використанні лише стандартних засобів оператор витрачає близько 12-16 годин.
- Групі стенографістів з 5 осіб для обробки запису знадобилося близько 40 хвилин, після чого створений файл було направлено на обробку коректору. Разом з корекцією розшифровка стенограми зайняла близько однієї години.

Проведені експерименти демонструють ефективність запропонованої системи порівняно з використанням лише традиційних засобів. Разом з іншими перевагами, серед яких відсутність необхідності адміністрування і впровадження, простота користування і зрозумілий інтерфейс користувача і якісна попередня обробка сигналу, запропонована система є досить ефективним засобом для автоматизації процесу створення і розшифровки стенограм засідань для невеликий і великих організацій, а також для індивідуальних користувачів.

6. Література

[1] *Система* распределенного компьютерного документирования устных выступлений и фонограмм речи Нестор // http://www.speechpro.ru/rus/products/doc-systems/nestor/

[2] *Комплекс* оперативного стенографирования “SRS Report 2002” // http://srs.kiev.ua/index.php?pg=2&lang=rus&tov=23

[3] *The meeting* recorder project // http://www.icsi.berkeley.edu/Speech/mr/mtgrcdr.html

[4] *Metze F., Jin Q., Fugen C., Laskowski K., Pan Y., Schultz T.* Issues in Meeting Transcription. – The ISL Meeting Transcription System // http://isl.ira.uka.de/fileadmin/publication-files/islMeetingSystem_icslp04.pdf

[5] *Yu H., Tomokiyo T., Wangand Z., Waibel A.* New Developments in Automatic Meeting Transcription // Proceedings of ICSLP2000, 2000. http://www.is.cs.cmu.edu/papers/speech/ICSLP2000/ICSLP2000-hua1.pdf

[6] *Hain T., Burget L., Dines J., Garau G., Karafiat M., Linkoln M., Vepa J., Wan V.* The AMI Meeting Transcription System: Progress and Performance, 2006. // http://www.cstr.ed.ac.uk/downloads/publications/2006/AMIasr.nist06.pdf

[7] *Yu H., Clark C., Malkin R., Waibel A.* Experiments in automatic meeting transcription using JRTK. // Acoustics, Speech and Signal Processing, 1998. Proceedings of the 1998 IEEE International Conference. – 1998. 12-15 May. – Vol.2. – P.921 – 92.