Адаптивний алгоритм визначення пауз у голосовому мовному сигналі для задач автоматизованого стенографування

Загваздін О. С.

Інститут кібернетики НАН України ім. Глушкова
відділ 115
alex.zagvazdin@gmail.com

1. Вступ

В рамках створення системи автоматизованого розподіленого стенографування [1, 2] однією з найважливіших задач була сегментація вхідного звукового сигналу на рівноцінні короткі сегменти так, щоби границі сегментів були пов'язані з паузами у мовному потоці. Задача сегментації сигналу складається в загальному сенсі з двох підзадач:

1) виділення пауз у мовному потоці

2) розбиття сигналу на сегменти визначеної тривалості.

Задача виділення пауз в мовному сигналі є важливою задачею не лише для сегментації сигналу, а й для задач очищення звукового сигналу від шумів, стиснення звукового сигналу в системах телекомунікації, розпізнавання мовних сигналів, тощо. В роботі запропоновано новий підхід до визначення пауз у мовному сигналі, що може бути застосовані для широкого класу задач.

Задача визначення пауз у мовному сигналі в реальних умовах ускладнюється тим, що рівень шуму у тривалому мовному сигналі не є стаціонарним, тобто співвідношення сигнал/шум може бути різним в різних ділянках сигналу. Як вказано у роботі Hirsch [3], «ця задача є дуже складною і в загальному сенсі не розв'язаною для реалістичних ситуацій зі змінним рівнем шуму». Проте існує досить велика кількість алгоритмів для визначення пауз у мовному потоці, кожен з яких має певні переваги у швидкодії, чутливості, точності чи обчислювальній складності. Ранні алгоритми базуються на відстані між коефіцієнтами лінійного передбачення Itakura [4], рівнях енергії, пітчі, і частоті зміни знаку амплітуди сигналу (zero crossing rate) [5], кепстральних властивостях [6], тощо. Деякі з цих алгоритмів на сьогодні успішно використовуються, зокрема у телекомунікаційних системах [7], проте, оскільки більшість з них базується на стаціонарно визначеному пороговому рівні, що відділяє ділянки звукового сигналу де є пауза, якість визначення пауз такими алгоритмами є не достатньо високою для сигналів зі змінним рівнем шуму і невисоким співвідношенням сигнал/шум.

Останнім часом було розроблено досить багато адаптивних алгоритмів, що можуть адаптуватися до змінного співвідношення сигнал/шум і можуть впевнено визначати паузи у мовному сигналі з низьким співвідношенням сигнал/шум. Так, в роботі [8] наводиться порівняльний аналіз трьох алгоритмів визначення пауз, що базуються на логарифмічній енергії сигналу і порогових значеннях для енергії ділянок сигналу з голосовою активністю і ділянок сигналу лише з шумом (пауз). Показано, що точність адаптивних алгоритмів, які обраховують порогові значення на кожному кроці є вищою, аніж алгоритму, де порогові значення задано заздалегідь, проте поведінка запропонованих алгоритмів для сигналів з реальним шумом не є достатньо проілюстрованою.

Один з достатньо точних підходів до визначення пауз пропонують так звані геометрично адаптивні методи визначення порогової енергії, що базуються на імовірнісному розподілі амплітуди сигналу , а також комбінації класичних методів визначення порогової енергії з геометрично адаптивними. Проте такі алгоритми мають достатньо високу обчислювальну складність, що ускладнює їх використання у системах реального часу.

В більшості класичних підходів енергія сигналу визначається логарифмічно за амплітудою сигналу чи за частотою зміни знаку функцією амплітуди сигналу, проте деякі автори пропонують використовувати четвертого [9] порядку для визначення енергії сигналу.

В даній роботі запропоновано алгоритм визначення пауз і сегментації голосового мовного сигналу, де в якості енергії сигналу використовується дисперсія (статистика другого порядку) амплітуди сигналу, а порогові значення для відокремлення ділянок сигналу з паузами обраховуються адаптивно на кожному кроці алгоритму.

2. Визначення пауз у голосовому мовному сигналі

Для визначення пауз у голосовому мовному сигналі по сигналу проходимо прямокутним вікном довжиною 100 мс. таким чином, щоб кожне наступне вікно починалося на середині попереднього. Тобто на кожному кроці алгоритму сигнал задається як:

$$s_k[i] = \begin{cases} s[i] & 0.0\,5D*k \le i < 0.1\,5D*k \\ 0, & else \end{cases} \quad (1)$$

Тут D - частота дискретизації сигналу в Гц, а k - номер кроку алгоритму.

Вважаємо, що на ділянці сигналу тривалістю 10 секунд має бути принаймні одна пауза. Отже на кожному кроці алгоритму інформація про попередні 10 с. звукового сигналу має бути врахована. Також вважаємо, що енергія сигналу у ділянці з паузою є меншою, ніж енергія сигналу у ділянках, де присутня голосова активність.

2.1. Енергія сигналу

Енергію сигналу визначимо як дисперсію (статистику другого порядку) амплітуди сигналу у заданому вікні:

$$E_{s_k} = \frac{1}{N}\sum_{i=1}^{N} s_k[i]^2 - \left(\frac{1}{N}\sum_{i=1}^{N} s_k[i]\right)^2 \quad (2)$$

Можливість використання другого статистичного моменту в якості міри енергії сигналу показано в [9], а отже використання дисперсії в якості міри енергії є розвитком підходу описаного в [9]. Дисперсія в якості міри енергії сигналу має схожі властивості до логарифмічної енергії і частоти зміни знаку функцією амплітуди, обчислювальна складність для алгоритму обрахунку дисперсії є невеликою і при високій частоті дискретизації кількість значень, що враховуються при обрахунку дисперсії може буди зменшена за рахунок ігнорування кожного другого значення амплітуди сигналу у вікні, що дозволяє зменшити обчислювальну складність алгоритму ще більше без суттєвого погіршення точності алгоритму. До того ж, виходячи зі статистичних властивостей дисперсії вона характеризує відмінність значень у вибірці, що для ділянок сигналу з голосом має бути вищою, ніж для ділянок де присутній лише шум (пауз).

2.2. Обрахування порогових значень

Як зазначається в роботі [10], при неперервному мовленні в середньому кожне 28 з 34 слів має перед собою паузу тривалістю >= 250 мс. Виходячи з експериментальних даних, встановлено, що середня тривалість слова складає 0.25 – 1 с. Виходячи з цих даних, можна стверджувати, що на ділянці сигналу тривалістю 10 с. може бути в середньому 6-10 пауз. До того ж, як вказано у [14] доросла людина в середньому робить 14 дихальних рухів на хвилину, що означає, що протягом 10 с. людина . робить принаймні два вдохи, що, в свою чергу, означає, що на цій ділянці буде принаймні 2 паузи для вдоху повітря.

Враховуючи те, що при використанні систем стенографування досить часто доводиться мати справу з мовленням, швидшим за середню швидкість мовлення, є безпечним припустити, що на ділянці сигналу тривалістю 10 с. Буде присутня принаймні одна пауза, тривалість якої буде достатньою для сегментації звукового сигналу.

Отже, перед початком роботи алгоритму аналізуємо перші 10 секунд мовного звукового сигналу. Для аналізу проходимо по сигналу вікнами, як описано в (1) і для кожного вікна обраховуємо енергію за формулою (2). Серед всіх вікон на даній ділянці знаходимо вікно, для якого значенння енергії мінімальне.

Експериментально було встановлено, що навіть у достатньо зашумленому голосовому сигналі енергія вокалізовнаих і невокалізованих ділянок відрізняється принаймні в 4-6 разів. Отже мінімальне значення енергії домножуємо на 2 і отримуємо стартове порогове значенння для роботи алгоритму:

$$E_t = 2\min(E_{s_k}) \quad (3)$$

Тут E_{s_k} - енергія сигналу для кожного з вікон на перші 10 с. звучання, тобто

$$k = 1,2,...,200 \quad (4)$$

Враховуючи, що довжина вікна – 100 мс і вікна перекриваються на 50%.

Далі на кожному кроці алгоритму порівнюємо значення енергії у вікні, що аналізується, порівнюється з пороговим значенням і, якщо значення енергії у вікні, що аналізується, менше, ніж порогове значення:

$$E_{s_k} < E_t, \qquad k = 1,2,...,20 * t \quad (5)$$

стверджуємо, що у даній ділянці сигналу є пауза, де t - тривалість сигналу в секундах.

Порогове значення уточнюється на кожному кроці алгоритму, враховуючи інформацію про попередні 10 с. звучання сигналу: на кожному кроці алгоритми запам'ятовуємо в масиві значення енергії для поточного вікна, і виключаємо з масиву найперше значення (що було включено в масив раніше за всі інші). Аналогічно першому кроку алгоритму, знаходимо мінімальне значення в масиві, домножуємо його на 2і таким чином отримаємо нове порогове значення.

3. Результати експерименту

Роботу алгоритму було перевірено на записах фонограм засідань вченої ради інституту кібернетики НАН України ім.. Глушкова з приводу захисту дисертаційних робіт. Такі записи характеризуються тим, що мовлення під час засідання є досить швидким, паузи – короткими, а їх кількість - порівняно невеликою. До того ж, співвідношення сигнал/шум в таких сигналах є невеликим за рахунок того, що в сигналі присутній сторонні шуми від аудиторії і шуми від побутових звукозаписуючих пристроїв, якими створюються такі фонограми, є також досить значними.

Було проведено два експерименти: один з використанням запропонованого адаптивного алгоритму, а інший – з використанням статичного порогового значення, що було експериментально підібрано перед початком роботи алгоритму

Результати експерименту наведено в таблиці 1.

Таблиця 1: Результати експерименту з перевірки алгоритму визначення пауз у голосовому мовному сигналі.

Алгоритм	Реальна кількість пауз	Коректно знайдені паузи	Некоректно знайдені паузи	Пропущені паузи
Адапт.	17	14	4	3
Станд.	17	13	7	4

Зміну порогового значення при використанні адаптивного алгоритму показано на Рис. 1, на Рис. 2, проілюстровано використання стандартного алгоритму зі статичним значенням порогу.

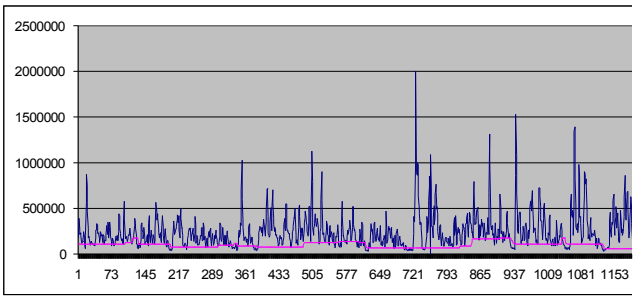

Рис. 1: Змінний поріг при використанні адаптивного алгоритму

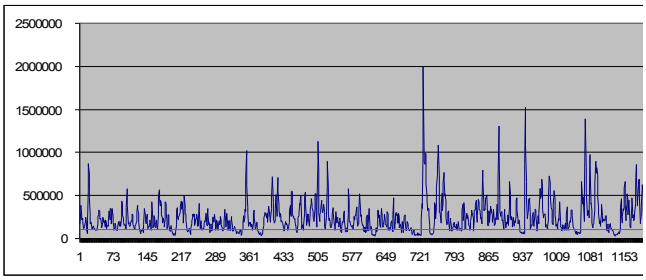

Рис. 2: Статичний поріг при використанні стандартного алгоритму

.Отже використання адаптивного алгоритму з перерахунком порогового значення дозволяє підвищити точність визначення пауз порівняно з алгоритмом, де використовується статичний поріг. Точність роботи алгоритму, зокрема зменшення кількості некоректно визначених пауз можна надалі покращити, ввевши обмеження, скажімо на тривалість паузи.

4. Висновки

Запропонований алгоритм визначення пауз у голосовому мовному сигналі дозволяє визначати паузи у неперервному мовному сигналі з точністю, достатньою для подальшої сегментації сигналу на короткі сегменти і обробки їх в системі автоматизованого стенографування. Завдяки використанню адаптивного порогу, алгоритм демонструє стабільно якісні результати в умовах нестаціонарного шуму.

5. Список літератури

[1] Ю.Г. Кривонос, Ю.В. Крак, А.В. Бармак, А.С. Загваздин. Информационная система распределенного компьютерного документирования речевых фонограмм заседаний. // Управляющие системы и машины, №3, 2008.

[2] Ю.Г. Кривонос, Ю.В. Крак, О.В. Бармак, О.С. Загваздін. Розподілене комп'ютерне документування голосових мовних фонограм. // Проблеми програмування. №2-3, 2008.

[3] H.G. Hirsch, "Estimation of noise spectrum and its application to SNR estimation and speech enhancement," Int. Comput. Sci. Inst., Berkeley, CA, Tech. Rep. TR-93-012, 1993.

[4] L. R. Rabiner and M. R. Sambur, "Voiced-unvoiced-silence detection using the Itakura LPC distance measure," in *Proc. Int. Conf. Acoust., Speech, Signal Processing*, May 1977, pp. 323–326.

[5] J. C. Junqua, B. Reaves, and B. Mak, "A study of endpoint detection algorithms in adverse conditions: Incidence on a DTW and HMM recognize," in *Proc. Eurospeech'91*, 1991, pp. 1371–1374.

[6] J. A. Haigh and J. S. Mason, "Robust voice activity detection using cepstral features," in *Proc. IEEE TENCON*, China, 1993, pp. 321–324.

[7] D. K. Freeman, G. Cosier, C. B. Southcott, and I. Boyd, "The voice activity detector for the Pan-European digital cellular mobile telephone service," in *Proc. Int. Conf. Acoust., Speech, Signal Processing*, Glasgow, U.K., May 1989, pp. 369–372.

[8] S. Gerven, F. Xie. A comparative study of speech detection methods. Proc. EUROSPEECH, 1997.

[9] E. Nemer, R. Goubran, and S. Mahmoud, "SNR estimation of speech signals using subbands and fourth-order statistics," *IEEE Signal Processing Lett.*, vol. 6, pp. 171–174, July 1999.

[10] A.W.Siegman. Of speech and time: temporal speech patterns in interpersonal contexts., 1979