УДК 004.934

Добрушкін Г. О., Данилов В. Я.

ПОРІВНЯННЯ ЯКОСТІ МЕЛ- ТА БАРК-ЧАСТОТНИХ КЕПСТРАЛЬНИХ КОЕФІЦІЄНТІВ ДЛЯ ПАРАМЕТРИЗАЦІЇ МОВНИХ СИГНАЛІВ

В даній статті представлені результати дослідження якості параметризації звукового сигналу кепстральними коефіцієнтами, частотна шкала яких викривлена трьома різними методами – використовуючи Барк-, Мел- і рівномірну шкали. Представлено теоретичний опис і емпіричні витоки Барк- і Мел-частотної шкали. Дослідження проведені на великій кількості тестових наборів з моделюванням різних варіантів зовнішніх шумових завад і різним рівнем сигнал/шум корисного сигналу. Результати роботи підтверджують те, що використання викривлених частотних шкал дозволяє досягти відчутно кращої якості параметризації сигналу в умовах низьких рівнів відношення сигнал/шум вхідного сигналу.

Ключові слова: *якість параметризації звукового сиглалу, кепстральні коефіцієнти, частотна шкала, Барк-, Мел- та равномірна шкала.*

В данной статье представлены результаты исследования качества параметризации звукового сигнала кепстральными коэффициентами, частотная шкала которых искривляется тремя разными методами – используя Барк-, Мел- и равномерную шкалу. Представлено теоретическое описание и эмпирические истоки Барк- и Мел-частотных шкал. Исследования проведены на большом количестве тестовых наборов с моделированием разных вариантов внешних шумовых помех и разных уровнем сигнал/шум полезного сигнала. Результаты работы подтверждают то, что использование частотных шкал искривления позволяет добиться ощутимо лучшего качества параметризации сигнала в условиях сильной зашумленности.

Ключевые слова: *качество параметризации звукового сигнала, кепстральные коэффициентами, частотная шкала, Барк-, Мел- и равномерную шкалы.*

This article investigates the quality of three different speech signal parameterization methods – cepstral coefficients, based on mel-frequency warp scale (MFC), bark-frequency warp scale (BFC) and on uniform scale (UFC). The theoretical description and empirical origins of Bark and Mel-frequency scales are presented. Investigations were carried out on a large number of test speech signals with wide variety of external noise and SNR level of useful signal were modeled. The results of the study confirm that the use of the frequency-warping scale dramatically increase recognition quality in heavy noised environment.

Key words: *quality of three different speech signal parameterization methods, cepstral coefficients, based on mel-frequency warp scale (MFC), bark-frequency warp scale (BFC) and on uniform scale (UFC).*

Вступ

Незважаючи на те, що в системах автоматичного розпізнання мовлення широко використовуються параметризація мовних сигналів, що ґрунтується на кепстральних коефіцієнтах, що використовують Мел-шкалу, в цій статті буде показано, що на справді Мел-шкала дає не дуже великі переваги в порівнянні з іншими частотно-деформаційними шкалами, що ґрунтуються на моделях людського сприйняття звуку. Виявилося, що подібна до MFCC функція, що ґрунтується на шкалі Барка дає аналогічні показники при розпізнаванні мови, як і MFCC. Якість MFCC і BFCC параметризації також порівнювалося з кепстральними коефіцієнтами на рівномірній шкалі (UFCC), що виявило те, що ні Мел ні Барк шкали не забезпечують значної переваги в порівнянні з рівномірною шкалою, якщо умови для тренувань і випробувань однакові.

Теоретичні засади

Мел-частотні кепстральні коефіцієнти (MFCC – Mel-frequency Cepstral Coefficients) широко використовуються в системах автоматичного розпізнавання мовлення (АРМ). Значення MFCC отримуються з модуля спектру (модуль значень швидкого перетворення Фур'є вхідного сигналу) із застосуванням банку фільтрів, які рівномірно розподілені на «викривленій», за відповідним законом, частотній шкалі. Далі

отриманий спектр зважується банками фільтрів, отримується набір значень, який логарифмується і потім декорелюється дискретним косинусним перетворенням (ДКП). Результатом останньої дії є вектор кепстральних коефіцієнтів. Описаний процес продемонстровано на рис. 1. Якщо для деформації частотної шкали використовується Мел-шкала, то отримані коефіцієнти називають Мел-кепстральними (MFCC), якщо Барк-шкала – Барк-кепстральними (BFCC), якщо частотна шкала не деформується – отримуємо UFCC (Uniform-frequency Cepstral Coefficients).

Рис. 1. Алгоритм отримання кепстральних коефіцієнтів

Порівнюючи вид трикутних вікон, якими згортається спектр вхідного сигналу на MFCC, BFCC, UFCC шкалах (рис. 2) стає очевидним, що Барк- і Мел- фільтри мають вузьку смугу пропускання на низьких частотах і значно більшу на великих, тоді як UFCC має постійний інтервал і ширину трикутних фільтрів на всьому частотному діапазоні. Тобто MFCC та BFCC актуалізують низькочастотну інформацію і усереднюють високочастотні складові сигналу, тоді як UFCC не надає переваги конкретним спектральним характеристикам сигналу.

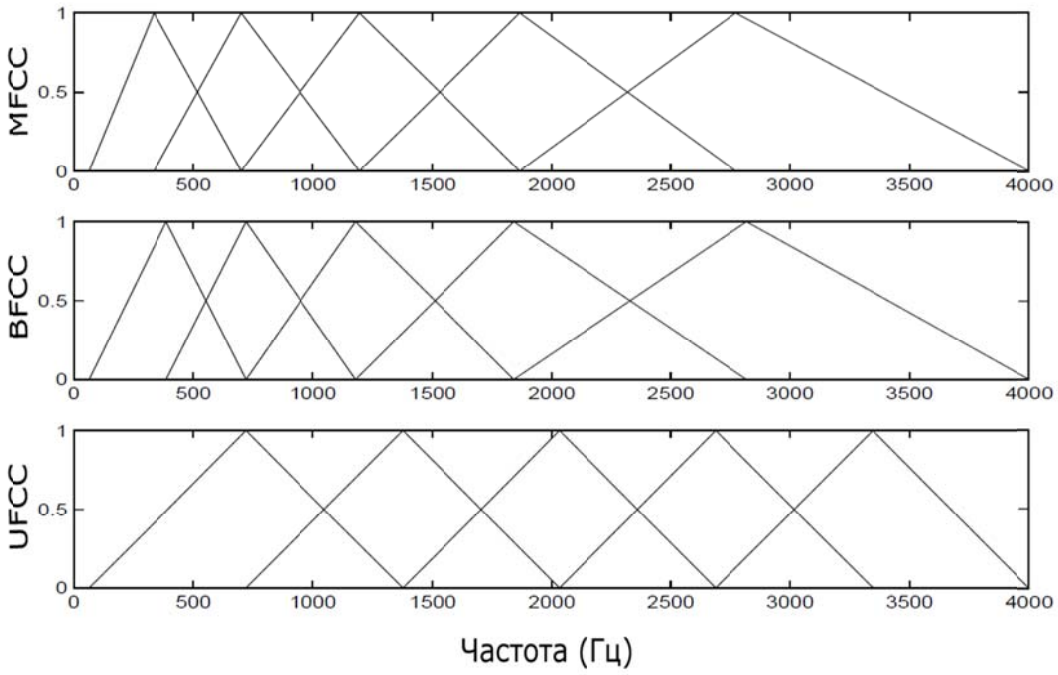

Рис. 2. Приклад банку із п’яти фільтрів на різних частото-деформуючих шкалах

Мел-шкала, це емпірична шкала, що ґрунтується на людському відчутті частоти звуку, була запропонована Стівенсом і Волкман в 1937 р. [1]. Шкала була отримана в результаті експериментів, в яких, випробовуваних просили скорегувати сигнал, який вони чують таким чином, щоб його висота стала в 2 рази нижчою. В результаті була отримана шкала, в якій 1000 Мел відповідає «висоті» звуку з частотою 1 кГц і подвоєння Мел створює відчуття сприйняття подвоєння висоти звуку. З емпіричних даних було отримано наступні аналітичні формули для розрахунку *mel*-шкали [2]:

$$mel(f) = 2595 \cdot \lg\left(1 + \frac{f}{700}\right).$$

Барк шкала є альтернативною емпіричною шкалою, що також ґрунтується на людському сприйнятті. Чіткість сприйняття голосу у людей починається зі спектрального аналізу, що виконується базальною

мембраною (БМ) в завитці вуха. Кожна точка на БМ може розглядатися в якості смугового фільтра з смугою пропускання, що дорівнює 1 критичній смузі або 1 Барк. Пропускна здатність ряду слухових фільтрів емпірично спостерігається і використані при розробці Барк шкали. Барк-шкала була розроблена Еберхардом Цвікером в 1961 році і названа на честь Генріха Баркгаузена, який запропонував першу модель суб'єктивного виміру гучності [3]. Функція, що задає Барк-шкалу зазвичай визначається таблично і є відображенням 24 критичних смуг слуху в числа з відрізку [1, 24], але існує декілька формул, що аналітично наближують емпіричні результати з достатньою точністю:

$$bark(f) \approx 6 \cdot \ln\left(\frac{f}{600} + \sqrt{1 + \left(\frac{f}{600}\right)^2}\right);$$

$$bark(f) \approx 13 \cdot atan\left(0{,}76 \cdot \frac{f}{1000}\right) + 3{,}5 \cdot atan\left[\left(\frac{f}{7500}\right)^2\right].$$

Графіки *mel*- та *bark*-перетворення наведені на рис. 3. На рис. 3 (б) суцільною лінією представлена друга формула *bark*(*f*).

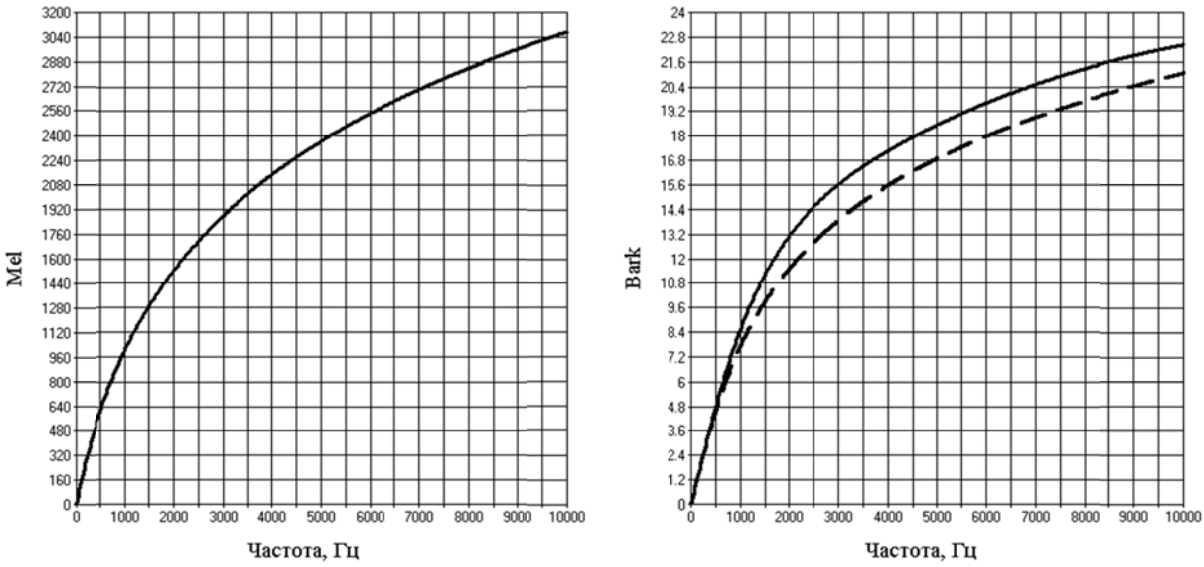

Рис. 3. Графіки аналітичних виразів, що використовуються для апроксимації *mel*- та *bark*-шкал

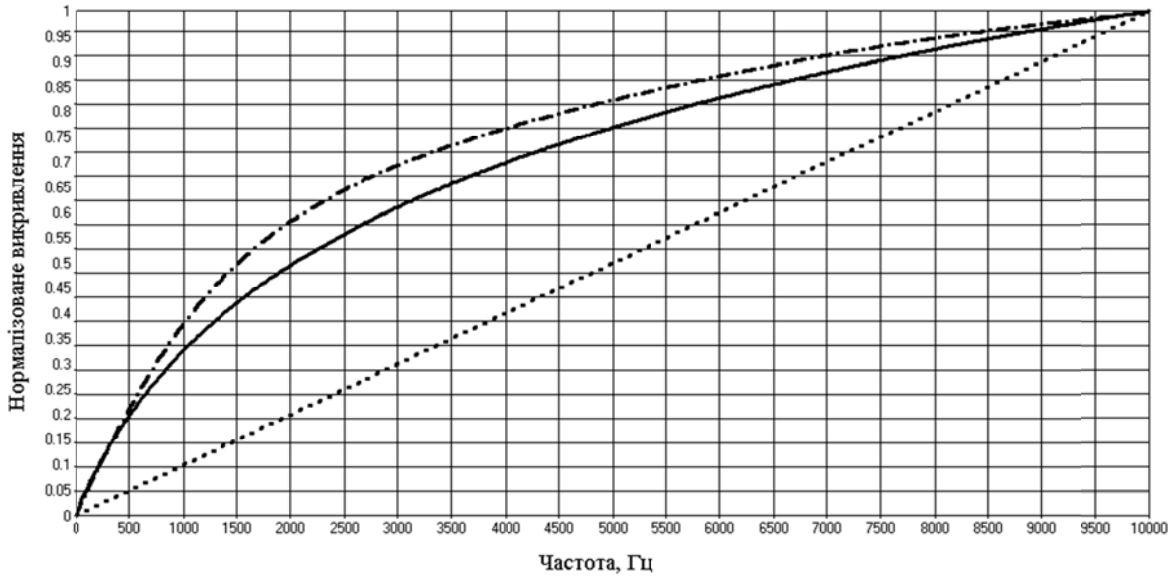

Рис. 4. Порівняльний графік нормалізованих *mel*-, *bark*- та рівномірного перетворення частотної шкали

В цій статті буде показано, що MFCC в якості алгоритму параметризації мовного сигналу дає такі ж результаті в розпізнаванні, як і BFCC. MFCC та BFCC також порівнюються з UFCC. Буде показано, що

шкала, що використовується для розташування фільтрів надає дуже малі переваги, особливо в однакових умовах для тренування и тестування.

Тренувальні і тестові набору

В якості тестових прикладів була використана база даних **NOIZEUS** [4] – спеціальна база даних, що використовується для дослідження алгоритмів покращення звуку і що являє собою сукупність мовних даних, на які накладено шуми, такі як шуми приміської *залізниці*, фонової *балаканини*, автомобільного *двигуна*, *виставкового зал*, *ресторану*, *вулиці*, *аеропорту* та станції *метро***.** Проводилося дикторонезалежне розпізнавання мовлення за чистих умов и з додаванням одного з зазначених видів шуму. Мовними даними в *NOIZEUS* є 30 речень англійської розмовної мови, вимовлених трьома чоловіками та трьома жінками (по 5 на кожного диктора), частота дискретизації при записі була 25 кГц, але задля додавання шуму була зменшена до 8 кГц. Речення відібрані такі, що мають низьку ймовірність передбачення наступного слова за контекстом речення і такі, що разом відбивають усі фонеми англійської мови.

Тренування проводилося в двох ситуаціях. В першій модель навчали чистими вимовами з БД, в другій – модель навчали вимовами, в які попередньо було додано шум. Перший навчальний набір мав 30 вимов, другий – 600. В другому випадку кожну вимову було подано в 20 різних варіаціях: 5 варіацій за рівнем сигнал/шум (чистий звук, 20 дБ, 15 дБ, 10 дБ, 5 дБ), кожна з яких ще мала 4 варіації за типом штучно доданого фонового шуму (*метро*, *балаканина*, *аеропорт*, *виставковий зал*).

Тестування системи виконувалось на наборі з 720 вимов и були однаковими для обох навчальних наборів. Тестовий набір складався з 24 груп по 30 вимов кожна. Групи наступні: 6 груп за відношенням сигнал/шум (чистий звук, 20 дБ, 15 дБ, 10 дБ, 5 дБ, 0 дБ), кожна з яких ще мала 4 варіації за типом штучно доданих фонових шумів (*метро*, *балаканина*, *аеропорт*, *виставковий зал*).

Інструментальний набір

Для тестування моделей використовувалася набір інструментів HTK (Hidden Markov Model Toolkit). Кожне слово, представлене в тестових реченнях змодельоване прихованою Марківською Моделлю (ПММ) з неперервною щільністю. Кожна ПММ має 18 явних станів з трьома Гаусівськими змішаним моделями для кожного стану.

Видобуток параметрів

Як було зазначено вище, вхідний сигнал поділяється на кадри, тривалістю 12.5 мс, з кожного з яких видобувається 36-мірний вектор параметрів. Кожний вектор параметрів складається з 12 кепстральних коефіцієнтів (за винятком , який не несе інформації про спектр сигналу, а лише інформацію про його енергію), 12 дельта-коефіцієнтів (перша похідна) і 12 подвійних дельта кепстральних характеристик (друга похідна). Так як c_0 виключено, то в експериментах, в векторі параметрів рівень енергії не використовувався.

Кількість трикутних смугових фільтрів в наборі дорівнювала 23. Координати точок фільтрів визначаються так, що кожна пара фільтрів перекривається на 50 % і на викривленій частотній шкалі кожен фільтр починається і закінчується в центрі сусідньої фільтра. Коли фільтри повертаються в шкалу герців вони зберігають трикутну форму, але змінюють своє розташування (рис. 2). Банк з 23-х фільтрів знаходиться в проміжку між 40 і 4000 Гц.

Оцінка моделей

Оцінка моделей проводилася за виразом з HTK для оцінки лінгвістичних моделей. HTK, оцінюючи лінгвістичну модель порівнює вихідну транскрипцію створену моделлю з вхідною транскрипцією. Порівняння виконується з використанням динамічного програмування, метою якого є найбільш вдале вирівнювання строк, що порівнюються з мінімізацією помилок «заміни», «вставки» та «видалення».

Власне точність розпізнавання розраховується за наступним виразом:

$$Точність = \frac{N - D - S - l}{N} \times 100\ \%,$$

де N – загальна кількість слів у виразі, D – кількість помилок типу «видалення», S – кількість помилок типу «заміна», l – кількість помилок типу «вставка».

Результати експериментів

Навчання за чистих умов

За умови навчання і тестування без наявності шумів всі три набори параметрів показали гарні результати (94,31 % – 96,32 %), але точність розпізнавання значно погіршується з введенням шумів у тестовий набір (56,26 % – 74,27 % при відношенні сигнал/шум 10 дБ). В цьому експерименті різниця в якості розпізнавання між Мел- і Барк- шкалами була незначною. Показники точності розпізнавання для обох варіантів (Мел, Барк) шкал і усіх чотирьох різновидів шуму накладаються один на одного (рис. 4). Для рівномірної шкали найбільший програш в акуратності розпізнавання досяг 8.27 % відносно точності найближчої шкали (BFCC, шум – *виставковий зал*, відношення сигнал/шум 10 дБ) (табл. 1).

Навчання з шумами

У цьому експерименті акуратність Мел- і Барк- шкал відрізнялися незначно (рис. 5). Точність розпізнавання для всіх трьох варіантів шкал зі зменшенням відношення сигнал/шум залишається великою, що і слід було очікувати, так як в цьому експерименті система вже навчалася на зашумлених зразках. На відміну від експерименту з навчанням за чистих умов, UFCC набагато ближче до інших шкал. Найбільше відставання було зафіксовано від Барк шкали при відношенні сигнал/шум у 15 дБ з фоновим шумом у вигляді *балаканини* – 7,04 % (табл. 1).

Таблиця 1.

Результати експериментів

	Навчання за чистих умов				Навчання з шумами			
	MFCC				**MFCC**			
	Метро	Балаканина	Аеропорт	Виставка	Метро	Балаканина	Аеропорт	Виставка
Чистий	95,57	95,87	95,13	95,14	90,38	89,78	90,54	89,21
20 дБ	94,01	90,34	93,65	93,62	94,42	92,26	93,26	92,47
15 дБ	89,43	87,28	88,12	89,73	93,18	91,11	91,81	91,59
10 дБ	69,21	73,12	63,38	63,38	90,27	89,03	90,64	89,61
5 дБ	36,67	50,35	30,28	29,94	81,16	82,70	76,34	76,54
0 дБ	18,41	22,01	16,18	15,63	47,57	49,28	35,48	41,27
Середнє	67,22	69,83	64,46	64,57	82,83	82,36	79,68	80,12
	BFCC				**BFCC**			
	Метро	Балаканина	Аеропорт	Виставка	Метро	Балаканина	Аеропорт	Виставка
Чистий	96,32	95,90	95,68	95,42	91,30	90,67	91,19	90,17
20 дБ	94,22	90,78	93,70	93,18	95,21	93,39	94,34	93,26
15 дБ	90,56	87,89	89,27	89,78	94,79	91,40	92,48	92,17
10 дБ	70,98	74,27	65,37	64,53	91,38	90,14	91,20	90,21
5 дБ	38,10	51,30	31,21	30,44	82,05	83,41	76,83	77,81
0 дБ	18,24	22,25	17,13	16,89	47,94	50,74	35,81	42,34
Середнє	68,07	70,40	65,39	65,04	83,78	83,29	80,31	80,99
	UFCC				**UFCC**			
	Метро	Балаканина	Аеропорт	Виставка	Метро	Балаканина	Аеропорт	Виставка
Чистий	95,68	95,11	95,32	94,31	89,84	89,54	89,17	89,24
20 дБ	92,71	88,24	92,27	91,05	93,37	86,47	92,21	91,84
15 дБ	86,38	81,05	83,38	81,61	91,29	84,36	90,28	90,43
10 дБ	63,42	67,00	60,81	56,26	86,17	84,27	88,61	88,21
5 дБ	30,57	43,65	27,47	24,17	76,41	78,55	78,57	73,90
0 дБ	20,36	21,38	18,61	17,34	41,58	46,38	36,38	35,74
Середнє	64,85	66,07	62,98	60,79	79,78	78,26	79,20	78,23

Висновки

Несподіваним виявився той факт, що кепстральні коефіцієнти, що ґрунтуються на Мел- і Барк-шкалах виявилися майже ідентичними в своїй здатності якісно чисельно представляти мовний сигнал. Але тим не менш, коли навчання і тестування проводиться за різних умов, наприклад, за відсутності і в присутності шумів відповідно, використання будь-якої з частото-деформуючих шкал виявилося значно кращими, аніж використання рівномірної шкали. Саме з такими умовами середовища стикаються розробники автоматичних систем розпізнавання мови при впровадженні своїх наробок у життя.

Коли і навчання і тестування проводиться за умов відсутності шумів взагалі, або присутності однакових шумів – жодна шкала не показала значної переваги над іншими.

ЛІТЕРАТУРА

1. Stevens Stanley Smith. A scale for the measurement of the psychological magnitude of pitch / Stanley Smith Stevens, John Volkman & Edwin Newman // Journal of the Acoustical Society of America. – 8 (3). – P. 185–190.
2. Beranek Leo L. Acoustic measurements. – New York : McGraw-Hill.
3. Zwicker E. Subdivision of the audible frequency range into critical bands // The Journal of the Acoustical Society of America. – 33. – Feb. – 1961.
4. http://www.utdallas.edu/~loizou/cimplants/quality_spcom07.pdf.

Стаття надійшла до редколегії 22.04.2011 р.