Известия Гомельского государственного университета
имени Ф. Скорины, № 6 (117), 2019

Информатика

УДК 004.934.2

Метод шумоочистки речевых сигналов на основе мел-частотных кепстральных коэффициентов с использованием фильтрации Калмана

С.М. Горошко, С.Н. Петров

Улучшение качества речи в системах автоматического распознавания с использованием фильтра Калмана является перспективной областью исследований. В подавляющем большинстве работ коэффициент линейного предсказания (*LPC)* используются в качестве параметрических векторов. Однако несколько исследований показали преимущество использования мел-частотных кепстральных коэффициентов (*MFCC*) в сравнении с *LPC* в системах распознавания речи. В данной статье проведен сравнительный анализ использования *LPC* и *MFCC* совместно с фильтром Калмана для определения акустических параметров зашумленного речевого сигнала в системах автоматического распознавания речи (*ASR*).
Ключевые слова: мел-частотные кепстральные коэффициенты, улучшение качества речи, извлечение характеристик, фильтр Калмана, автоматическое распознавание речи.

Improving speech quality in automatic recognition systems using the Kalman filter is a promising area of research. The vast majority of work done in this area uses linear prediction coefficients (LPC) as a parametric vector. However, several studies have shown the advantage of using mel-frequency cepstral coefficients (MFCC) in comparison with LPC in speech recognition systems. This article presents a comparative analysis of the use of LPC and MFCC in conjunction with the Kalman filter to determine the acoustic parameters of a noisy speech signal in automatic speech recognition (ASR) systems.
Keywords: mel-frequency cepstral coefficients, speech enhancement, feature extraction, Kalman filter, automatic speech recognition.

Введение. Автоматическое распознавание речи – это технология, обеспечивающая процесс преобразования речевых сигналов, произнесенных людьми, в текстовый поток данных [1]. Улучшение фундаментальных подходов и новых разработок привело к эволюции в области автоматического распознавания речи (*ASR*).

Традиционные системы распознавания обучаются речевым сигналам, записанным в чистой и бесшумной среде. Но в реальных условиях качество распознавания зависит от среды, через которую проходит акустический сигнал, что приводит к ухудшению функции распознавателей в шумных средах. Повышение разборчивости речи – это процесс цифровой обработки сигналов, который использует статистический характер речевых и шумовых сигналов для коррекции зашумленного речевого сигнала и получения высококачественного речевого сигнала. Обзор технологий шумоочистки речевых сигналов показал, что метод фильтрации с использованием фильтра Калмана дает превосходные результаты [2]. В данной работе фильтр Калмана использовался для уменьшения негативного воздействия случайных шумов в извлеченных параметрах речевых признаков.

С развитием вычислительной техники были разработаны различные методы извлечения полезных параметров из речевого сигнала [3]. В статье сравнивается эффективность метода *MFCC* и *LPC* наряду с использованием фильтра Калмана при подавлении шума. Реализация во временной области речевого сигнала, хотя и содержит в себе всю акустическую информацию, с трудом поддается анализу. Выделение функций позволяет выявить определенную статистически значимую информацию из речевого сигнала при использовании относительно небольшого числа параметров. Основное внимание уделено разработке фильтра с помощью полученных параметров, частотная характеристика которых эквивалентна спектральной плотности мощности речевого сигнала.

Математическая модель алгоритма обработки зашумленных речевых сигналов. Линейное предсказание – вычислительная процедура, позволяющая по некоторому набору предшествующих отсчётов цифрового сигнала предсказать текущий отсчёт. Формирование речевого сигнала можно упрощенно представить в виде БИХ-фильтра. Суть метода состоит в том, чтобы выразить каждый кадр как смесь предыдущих кадров. Для сглаживания переходов в речевом спектре каждые 3 мс синтезатор обновляет параметры *LPC*, проводя интерполяцию между параметрами предыдущего и следующего кадра [4]. Передаточная функция изменяющегося во времени фильтра задается уравнением 1:

$$H(z) = \frac{G}{1 + \sum a_p z^{-p}}, \tag{1}$$

где G – коэффициент усиления, a_p – коэффициенты *LPC* фильтра, p – порядок фильтрации.

Алгоритм вычисления коэффициентов представлен на рисунке 1.

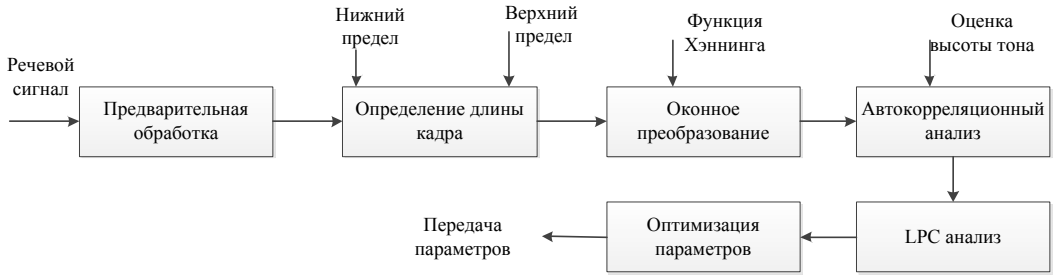

Рисунок 1 – Вычисление параметров *LPC*

Мел-частотные кепстральные коэффициенты (*MFCC*) были представлены в речевом анализе Дэвисом и Мермельштейном в 1980-х гг. [5]. Использование шкалы мела было обусловлено тем фактом, что люди могут различать небольшие изменения высоты тона на более низких частотах. Функции *MFCC* соответствуют энергиям банка логарифмических фильтров с использованием шкалы мела. Частота в масштабе мела определяется уравнением 2:

$$Mel(f) = 2595 \log(1 + \frac{f}{700}). \tag{2}$$

Векторы *MFCC* составляют около 39 векторов, включая 13 статических коэффициентов и 26 динамических коэффициентов, которые определяются разностью 1-го и 2-го порядка статических коэффициентов. Динамические коэффициенты фиксируют изменения коэффициентов во времени и часто не используются для распознавания речи. Алгоритм вычисления *MFCC* показан на рисунке 2.

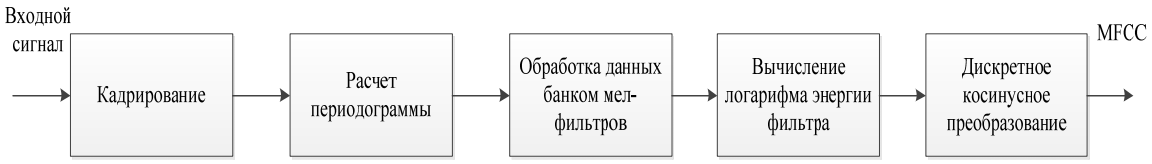

Рисунок 2 – Вычисление *MFCC*

Фильтр Калмана представляет собой математическую процедуру, разработанную Рудольфом Эмилем Калманом. Это алгоритм оптимальной рекурсивной обработки данных с использованием механизма коррекции предсказания. Алгоритм прогнозирует будущее состояние, используя предварительные аппроксимации, а затем корректирует дополнительный член, который пропорционален расчетной ошибке.

В случае обработки речи [1] принятый сигнал состоит из полезного сигнала и фонового шума. Пусть полученный сигнал задается уравнением 3:

$$Z_n = x_n + \nu_n, \tag{3}$$

где Z_n – полученный сигнал, x_n – входной речевой сигнал, ν_n – аддитивный фоновый шум.

Входной вектор моделируется с использованием авторегрессионной модели, заданной уравнением 4:

$$x_n = \sum_{k=1}^{m} a_k x_{(n-k)} + w_n. \tag{4}$$

Вышеприведенное уравнение может быть записано в форме состояния-пространства, как указано в уравнении 5:

$$x_{n+1} = A_{n+1,n} x_n + B.w_{n+1}, \tag{5}$$

где w_{n+1} – шум процесса, $A_{n+1,n}$ – транспонированная матрица состояний, представлена как

$$A_{n+1,n} = \begin{bmatrix} a_1 & -a_2 & \dots & -a_1 & -a_1 \\ 1 & 0 & \dots & \dots & 0 \\ 0 & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \dots & 0 & 1 & 0 \end{bmatrix}.$$

Выходной вектор можно задать как уравнение 6:

$$Z_n = H.x_n + \nu_n, \tag{6}$$

где H – состояние выходного вектора, ν_n – внешний шум.

Результаты экспериментов. В данной работе проводилось моделирование с помощью *MATLAB* 2017*a*. Речевые сигналы, необходимые для анализа, были отобраны из речевой базы данных *NOIZEUS* [6]. База данных состоит из 30 предложений *IEEE* sentence database в не зашумленной форме, а также в зашумленной форме с различными шумами реального мира с определенными уровнями отношения сигнал/шум (*SNR)* 0, 5, 10 и 15 дБ. Частота дискретизации составляет 8 кГц.

Целью метода выделения признаков является воссоздание спектральных характеристик мощности речевого сигнала. Фильтр, разработанный с использованием параметров *LPC*, имеет частотную характеристику, эквивалентную спектральной огибающей мощности кадра речевого сигнала (рисунок 3).

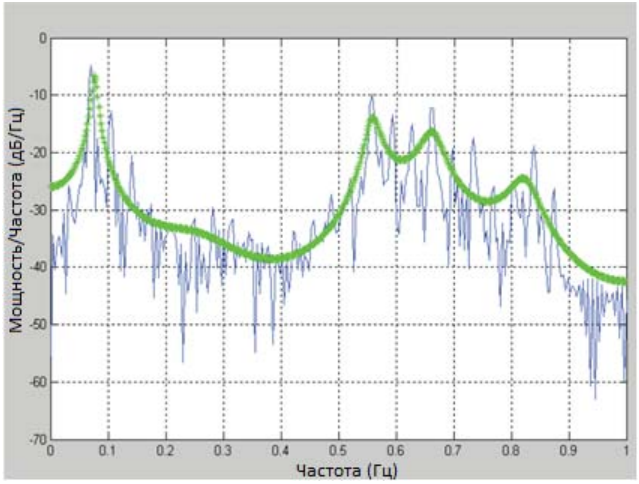

Рисунок 3 – Частотная характеристика фильтра *LPC* для первого входного кадра

Целью работы является вычисление параметров *LPC* из зашумленного речевого кадра, уточнение этих параметров с помощью фильтра Калмана, а затем восстановление речевого кадра из оценочных параметров. На рисунке 4 показан график зашумленного речевого кадра (1) и речевого кадра на выходе фильтра (2), а также график спектральной плотности мощности не зашумленного речевого кадра (3). Из графика можно сделать вывод, что при низкой энергии *LPC* оцениваемые параметры аналогичны параметрам зашумленного кадра. Это показывает, что *LPC* не может различать невокализованную речь (с низкой энергией) и шумы.

Тот же процесс извлечения функций *MFCC* из искаженного шумами сигнала и уточнения этих параметров с помощью фильтра Калмана дает лучшие результаты, чем *LPC* для большего уровня шумов (0 дБ и 5 дБ).

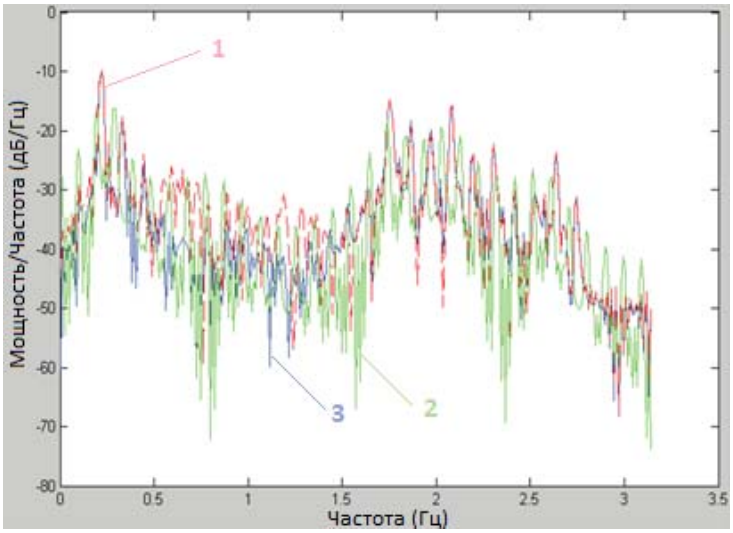

Рисунок 4 – Спектральная плотность мощности тестовых сигналов

Для получения результатов использовался речевой сигнал «sp02» из базы данных с наложением трех типов помех, а именно: шум аэропорта, шум ж/д станции и шум ресторана для 4 разных уровней (0,5, 10, 15 дБ). Параметры получены из чистого речевого сигнала и зашумленных модификаций для получения значений отношения сигнал/шум. Результаты моделирования приведены в таблице 1.

Таблица 1 – Результаты моделирования

Тип шума	SNR зашумленного сигнала (дБ)	SNR параметров			
		LPC		MFCC	
		До фильтрации	После фильтрации	До фильтрации	После фильтрации
Шум самолета	0	6,6252	6,8338	11,0347	**14,1640**
	5	10,9283	11,0109	12,3598	**13,5350**
	10	11,4412	11,6609	14,9375	**15,3048**
	15	20,1894	**20,3872**	22,0625	16,9968
Шум ж/д станции	0	7,0742	7,2136	9,0765	**11,3361**
	5	12,3574	12,5487	14,1404	**15,4818**
	10	13,7955	**13,8855**	19,2336	16,8116
	15	27,9943	**28,9882**	25,9601	17,8780
Шум помещения	0	8,3061	8,5388	12,2778	**14,8871**
	5	12,9780	12,9901	16,8215	**16,9546**
	10	26,8417	**26,9910**	25,9173	17,8637
	15	28,6348	**29,2603**	28,0637	18,1613

Параметры, которые были улучшены после фильтрации, выделены жирным шрифтом. На основе этих данных можно сделать вывод, что фильтр Калмана улучшает параметры *MFCC* значительно лучше, чем *LPC*, в более шумных условиях. Следует отметить, что *LPC* демонстрирует лучшие результаты при очень низком уровне шумов. В точке, близкой к 11–12 дБ, *MFCC* и *LPC* дают приблизительно одинаковые результаты.

Заключение. Получены практические результаты для параметрических векторов, основанных на мел-частотных кепстральных коэффициентах (*MFCC*) и коэффициентах линейного предсказания (*LPC*), и соответствующие оценки эффективности работы системы для различных параметров модели. В статье показано, что *MFCC* в сочетании с фильтрацией Калмана дает более точные результаты при распознавании речевого сигнала, однако при очень низком уровне шумов *LPC* предпочтительней.

Литература

1. Сорокин, В.Н. Распознавание личности по голосу: аналитический обзор / В.Н. Сорокин, В.В. Вьюгин, А.А. Тананыкин // Информационные процессы. – 2012. – Т. 12, № 1. – С. 1–30.

2. Джалолов, У.Х. Регуляризация задачи идентификации объекта в условиях зашумленности полезного сигнала / У.Х. Джалолов, Р.М. Бандишоева, У.А. Турсунбадалов // Вестник Таджикского технического университета. – 2016. – № 1 (33). – С. 20–26.

3. Спажакин, Ю.Г. Современные средства проектирования систем голосовой биометрии / Ю.Г. Спажакин // Методы и устройства передачи и обработки информации. – 2016. – № 10 – С. 69–79.

4. Козырев, М.О. Оконные функции и преобразование Фурье / М.О. Козырев, М.Ю. Орлов // Инновационные научные исследования: теория, методология, практика : сб. статей IX Международной научно-практической конференции. – Пенза, 2017 – С. 21–25.

5. Гришунов, С.С.Основные математические методы выделения речевых особенностей в системах распознавания диктора / С.С. Гришунов, Ю.С. Белов // Наука, техника и образование. – 2015. – № 3 (3). – С. 53–58.

6. NOIZEUS: A noisy speech corpus for evaluation of speech enhancement algorithms [Electronic resource] : UN Bibliogr. Inform. System. – Mode of access : https://ecs.utdallas.edu/loizou/speech/noizeus/. – Date of access : 24.06.2019.

Белорусский государственный университет
информатики и радиоэлектроники

Поступила в редакцию 09.07.2019