КЕПСТРАЛЬНЫЕ КОЭФФИЦИЕНТЫ КАК НЕОБХОДИМАЯ ХАРАКТЕРИСТИКА ПРОЦЕССА СОЗДАНИЯ СИСТЕМЫ ИМИТАЦИИ ГОЛОСА ЧЕЛОВЕКА С ПОМОЩЬЮ МЕТОДОВ ГЛУБОКОГО ОБУЧЕНИЯ

Маркина Ю.Ю., Белов Ю.С.

Калужский филиал федерального государственного бюджетного образовательного учреждения высшего образования «Московский государственный технический университет имени Н.Э. Баумана (национальный исследовательский университет)», г. Калуга (248000, г. Калуга, ул. Баженова, д. 2.), e-mail: yakunina-julia25@yandex.ru

В большинстве случаев речевой сигнал не подается напрямую в систему имитации или системы по преобразованию голоса. Обычно речевой сигнал представляется в виде набора компактных характеристик для дальнейшего удобного управления им. Для оптимального выбора варианта представления спектральных данных в процессе создания модели системы имитации голоса человека требуется выполнение ряда условий, в который входят качественная передача индивидуальных особенности речи диктора как личности, соответствие в достаточной степени спектральной огибающей и обладание способностью в нее преобразовываться, наличие качественных свойств интерполяции и возможности гибкой модификации. Кепстральные коэффициенты в наилучшей степени отвечают данным условиям. Также необходимо прибегнуть к одномерному представлению спектральных коэффициентов с целью уменьшения занимаемого объема памяти. Частным случаем кепстральных коэффициентов являются Мел-кепстральные коэффициенты, они сосредоточены на частотах, которые имеют большое значение для человеческой речи и слуха. Использование кепстрального анализа широко распространено в задачах, связанных с обработкой голоса человека. Выбор кепстральных коэффициентов позволяет достигать лучшей производительности и обеспечивает качественное квантование, интерполяцию и представление формантной структуры.

Ключевые слова: имитация голоса человека, обработка речи, спектральные данные, кепстральные коэффициенты, Мел-кепстральные коэффициенты, кепструм.

CEPSTRAL COEFFICIENTS AS A NECESSARY CHARACTERISTIC OF THE PROCESS OF CREATION OF THE HUMAN VOICE IMITATION SYSTEM WITH THE HELP OF DEEP-LEARNING METHODS

Markina Yu.Yu., Belov Yu.S.

Kaluzhskij filial federal'nogo gosudarstvennogo byudzhetnogo obrazovatel'nogo uchrezhdenija vysshego obrazovanija «Moskovskij gosudarstvennyj tehnicheskij universitet imeni N.E. Baumana (nacional'nyj issledovatel'skij universitet)» , Kaluga (248000, Kaluga, Bazhenova st, 2.), e-mail: yakunina-julia25@yandex.ru

In most cases, the speech signal is not fed directly to the voice imitation or voice conversion system, but is presented as a set of compact characteristics for the further convenient manipulation of it. For optimal choice of the spectral data representation in the process of creating a model of the human voice imitation system, a number of conditions are required, which include a qualitative transfer of the individual features of the speaker's speech as a person, the adequacy of the spectral envelope and the ability to transform into it, the presence of qualitative interpolation properties and the possibility of flexible modification. It is also necessary to resort to a one-dimensional representation of the spectral coefficients to reduce the occupied memory capacity. A special case of cepstral coefficients is Mel-cepstral coefficients, they are concentrated on frequencies that are of great importance for human speech and hearing. The use of cepstral analysis is widespread in tasks related to speech processing. The choice of cepstral coefficients allows to achieve better performance and provides qualitative quantization, interpolation and representation of the formant structure.

Key words: imitation of human voice, speech processing, spectral data, cepstral coefficients, Mel-cepstral coefficients, cepstrum.

Речь каждого человека является неповторимой и имеет целый ряд индивидуальных особенностей. Голос человека столь же уникален, как строение сетчатки глаза или отпечатки пальцев. По голосу человека можно судить о его характере. Также известно, что общее впечатление о человеке наполовину зависит от мимики, на треть - от голоса и только лишь на малую часть - от того, что этот человек говорит.

Индивидуальные особенности голоса говорящего можно использовать не только в задачах идентификации, но и для определения настроения человека, борьбы с телефонными мошенниками и т.д. Круг прикладных задач этим не ограничивается. Так, например, существует зависимость между частотами среднего тона и гендерной принадлежностью - мужчины говорят на частоте 85-200Hz, а женщины — 160-340Hz. Таким образом, можно придумать множество прикладных сценариев, где выделение индивидуальных особенностей говорящего будет необходимо.

Обращаясь к проблеме распознавания речи, прежде всего, необходимо определить, что именно нужно сравнивать. Непосредственное сравнение звуковых сигналов во временной области является процессом долгим и неэффективным. Спектрограммы – более быстрый способ, но не намного эффективнее. Поиски максимально рационального представления приводят к кепстральным коэффициентам, которые часто используются в качестве характеристики речевых сигналов.

Как только слово выделяется из потока входных данных, начинается этап процесса выделение необходимых характеристик. В этом случае могут применяться различные методики, например методика нахождения мел-кепстральных коэффициентов или коэффициентов линейного предсказания. Основная задача на данном этапе - выделение неких параметров сигнала, причем число этих параметров должно быть минимально, чтобы ускорить сравнение с наборами параметров из библиотеки, и в то же время данные параметры должны быть такими, чтобы по ним можно было достаточно точно определить конкретное слово.

Таким образом, можно сделать вывод, что в задачах, связанных с обработкой речи, сигнал не подается напрямую в систему, а представляет собой набор компактных характеристик для наилучшего процесса их манипуляции. Более конкретно, если обратиться к рисунку 1, можно заметить, что для эффективного представления сигнала этап выделения ключевых характеристик речи присутствует как на фазе обучения, так и на фазе выполнения.

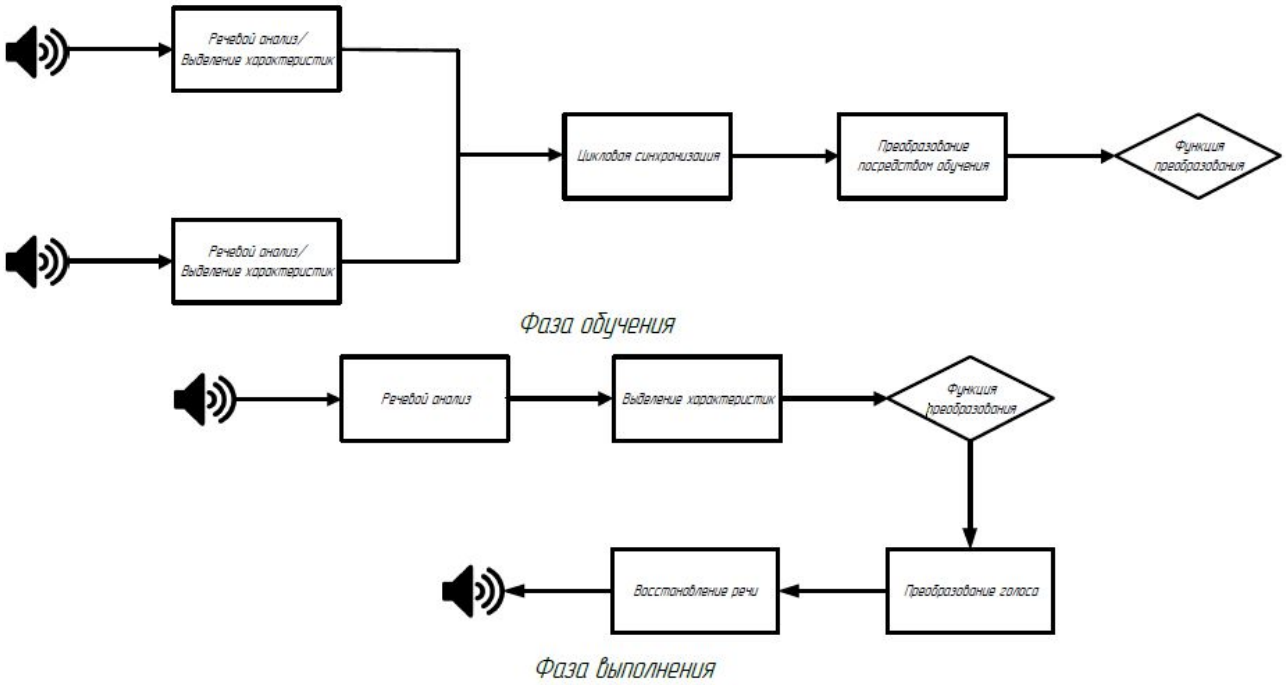

Рисунок 1 – Типовая система преобразования голоса

Обработка спектральных характеристик речи является следствием управления просодией. Однако существует множество представлений спектральных характеристик. Для процесса статистического параметрического моделирования, характеристики должны отвечать следующим требованиям [6]:

1. качественно передавать индивидуальные особенности речи диктора как личности;
2. в достаточной степени соответствовать спектральной огибающей и обладать способностью в нее преобразовываться;
3. обладать качественными свойствами интерполяции и возможностью гибкой модификации.

Спектральные характеристики являются многомерными, поэтому занимают значительную часть памяти при работе с большими объемами данных. Поэтому необходимо прибегнуть к наиболее эффективному одномерному представлению спектра, который может быть использован в дальнейшем на практике [7]. Поиски наиболее рационального представления приводят к кепстральным коэффициентам, которые часто используются в качестве характеристики речевых сигналов.

КЕПСТРАЛЬНЫЕ КОЭФФИЦИЕНТЫ

В процессе эволюции живые существа, обладающие сенсорными системами, развивались по принципу: «различать, для того чтобы выжить». Слуховой аппарат человека как сенсорный анализатор имеет способность обеспечивать различение звуков по их частотному составу. Однако реакция на звуковой стимул должна быть быстрой, а значит, обработка сигналов в ухе и нервной системе должна выполняться за небольшое время.

Требования высокой частотной и временной различительной способности анализатора противоречивы, но результатом эволюции было оптимальное сочетание этих показателей.

Органы слуха человека обладают свойством частотного маскирования, где под маскированием понимают ситуацию, при которой нормально слышимый звук накрывается другим громким звуком с близкой частотой. Данная характеристика зависит от частоты сигнала и варьируется от 100 Гц для низких слышимых частот до более 4000 Гц для высоких частот. Следовательно, область слышимых частот можно разделить на несколько критических полос (принято деление на 24 критические полосы), которые обозначают падение чувствительности уха для более высоких частот.

Можно считать критические полосы еще одной характеристикой звука, подобной его частоте. Однако, в отличие от частоты, которая абсолютна и не зависит от органов слуха, критические полосы определяются в соответствии со слуховым восприятием. В итоге они образуют некоторые меры восприятия частот, для которых введены единицы измерения - барк и мел.

Использование кепстрального анализа широко распространено в задачах, связанных с обработкой речи. Большинство современных автоматических систем синтеза и распознавания речи сосредотачивают усилия на извлечении частотной характеристики речевого тракта человека, отбрасывая при этом характеристики сигнала возбуждения. Это объяснено тем, что коэффициенты первой модели обеспечивают лучшую разделимость звуков. Для отделения сигнала возбуждения от сигнала речевого тракта прибегают к кепстральному анализу.

Модель кепстральных коэффициентов, как для минимумов, так и для максимумов оптимально соответствует спектральной огибающей - важной характеристике синтеза речи. Частный случай кепстральных коэффициентов - Мел-кепстральные коэффициенты, представляющие собой спектральную огибающую с коэффициентами, расположенными друг от друга на расстоянии по шкале Мела, которые сосредоточены на частотах, имеющих большое значение для человеческой речи и слуха [1]. Их использование в задаче описания характеристик фонемы обусловлено прежде всего удобством практического применения.

Мел-кепстральные коэффициенты обладают повышенной помехоустойчивостью и позволяют принимать достоверные решения на относительно коротких интервалах анализа речи. Основной идеей метода Мел-кепстральных коэффициентов является максимальное приближение информации поступающей на слуховой анализатор мозга человека. Признаки, построенные на основе Мел-кепстральных коэффициентов, учитывают психоакустические принципы восприятия речи, поскольку используют мел-шкалу, связанную с критическими полосами слуха.

Необходимо понимать значение понятий мела и кепстра. Мел - это единица высоты звука, которая основана на восприятии этого звука органами слуха человека или, другими словами, своеобразное представление энергии спектра сигнала, которое обычно является вектором из тринадцати вещественных чисел. Кепстр (cepstrum) - в свою очередь, это результат дискретного косинусного преобразования от логарифма амплитудного спектра сигнала.

Для того чтобы найти энергию сигнала, вектор спектра сигнала перемножается с функцией окна, в результате чего получается вектор коэффициентов. Если их возвести в квадрат, представить в виде логарифма и получить из них кепстральные коэффициенты, то получаются искомые Мел-коэффициенты. Кепстральные коэффициенты можно получить как с помощью Фурье-преобразования, так и с помощью дискретного косинусоидального преобразования [6]. Дискретное косинусоидальное преобразование применяется для получения кепстральных коэффициентов, оно сжимает полученные результаты, повышает вклад первых коэффициентов и понижает вклад последних.

Плюсы использования кепстральных коэффициентов заключаются в следующем:

- Используется спектр сигнала (то есть разложение по базису ортогональных косинусоидальных или синусоидальных функций), что позволяет учитывать волновую «природу» сигнала при дальнейшем анализе;
- Спектр проецируется на специальную Мел-шкалу, позволяя выделить наиболее значимые для восприятия человеком частоты;
- Количество вычисляемых коэффициентов может быть ограничено любым значением (например, 12), что позволяет «сжать» фрейм и, как следствие, количество обрабатываемой информации;

В работе [5] был предложен унифицированный подход к спектральному анализу речи, позволяющий вычислять различные наборы параметров. В эти параметры, помимо прочих, входят линейное предсказание и Мел-кепстральный анализ. За счет изменения параметров α и γ, есть возможность выбирать между доступными параметрами.

Кепстр $c(m)$ действительной последовательности $x(n)$ определяется как обратное преобразование Фурье логарифмического спектра, в то время как Мел-обобщенные кепстральные коэффициенты $c_{\alpha,\gamma}(m)$ определяются как обратное преобразование Фурье обобщенного логарифмического спектра, рассчитанное по деформированной частотной шкале $\beta_{\alpha,\gamma}(\omega)$:

$$s_\gamma\left(X\left(e^{j\omega}\right)\right) = \sum_{m=-\infty}^{\infty} c_{\alpha,\gamma}(m) e^{-j\,\beta_{\alpha,\gamma}(\omega)}, \quad (1)$$

где $s_\gamma(\omega)$ - обобщенная логарифмическая функция и $X(e^{j\omega})$ – преобразование Фурье для $x(n)$. Обобщенная логарифмическая функция определяется следующим образом:

$$s_\gamma(\omega) = \begin{cases} \frac{\omega^\gamma - 1}{\gamma} \\ \log\omega\,, \ \gamma = 0 \end{cases}, 0 < |\gamma| \leq 1, \quad (2)$$

а деформированная шкала частот $\beta_\alpha(\omega)$ рассчитывается как фазовая характеристика пропускающей системы [2]:

$$\psi_\alpha(z) = \frac{z^{-1}-\alpha}{1-\alpha z^{-1}} | z = e^{j\omega} = e^{-j\,\beta_\alpha(\omega)}, |\gamma| < 1, \quad (3)$$

где $\beta_\alpha(\omega) = \tan^{-1}\frac{(1-\alpha^2)\sin\omega}{(1+\alpha^2)\cos\omega - 2\alpha}$. (4)

В работе [4] допускается, что спектр речи $H(e^{j\omega})$ может быть смоделирован $M+1$ Мел-обобщенными кепстральными коэффициентами следующим образом:

$$H(z) = s_\gamma^{-1}(\sum_{m=0}^{M} c_{\alpha,\gamma}(m)\,\psi_\alpha^m(z)) \quad (5)$$

Выбрав $|\alpha| < 1$ и $\gamma = 0$, можно получить Мел-кепстральную функцию (MCEP):

$$exp\sum_{m=0}^{M} c_{\alpha,\gamma}(m)\,\psi_\alpha^m(z) = KD(z), \gamma = 0, \quad (6)$$

где

$$K = \exp\left(c_{\alpha,\gamma}(0)\right), \gamma = 0, \quad (7)$$

$$D(z) = exp\sum_{m=1}^{M} c_{\alpha,\gamma}(m)\,\psi_\alpha^m(z), \gamma = 0, \quad (8)$$

а $c = [c_{\alpha,\gamma}(0), c_{\alpha,\gamma}(1), \dots, c_{\alpha,\gamma}(M)]^T$ – вектор Мел-кепстральных коэффициентов со специальным коэффициентом $c_{\alpha,\gamma}(0)$, обычно называемым энергетической компонентой, так как он соответствует средней логарифмической мощности кадра [4]. Вычислительная сложность алгоритма кепстрального преобразования при использовании быстрого преобразования Фурье приближенно равна $O(N + 2N\log_2 N)$.

Мел-обобщенные кепстральные коэффициенты широко используются в проектах, связанных с преобразования голоса, позволяют достигать лучшей производительности по сравнению с другими вариантами и обеспечивают качественное квантование, интерполяцию и представление формантной структуры. Их использование также позволяет избегать создания артефактов в процессе синтеза речи[3].

Список литературы

1. Нифонтов С.В., Белов Ю.С. Применение скрытых марковских моделей в текстонезависимых системах идентификации пользователей по голосу// Электронный журнал: наука, техника и образование – №2, 2016. – с.116-124. [Электронный ресурс]. URL: https://elibrary.ru/item.asp?id=26847531 (дата обращения 17.09.2017)

2. Гришунов С.С., Белов Ю.С. Основные математические методы выделения речевых особенностей в системах распознавания диктора// Электронный журнал: наука, техника и образование – №3, 2015. – с.53-58. [Электронный ресурс]. URL: https://elibrary.ru/item.asp?id=26096853 (дата обращения 17.09.2017)

3. Белов Ю.С., Либеров Р.В. Подходы и проблемы распознавания личности по голосу// Электронный журнал: наука, техника и образование – №3, 2015. – с.68-77. [Электронный ресурс]. URL: https://elibrary.ru/item.asp?id=26096855 (дата обращения 17.09.2017)

4. Белов Ю.С., Нифонтов С.В., Азаренко К.А. Применение вейвлет-фильтрации для шумоподавления в речевых сигналах// Электронный журнал: фундаментальные исследования – №3, 2015. – с.29-33. [Электронный ресурс]. URL: https://www.fundamental-research.ru/ru/article/view?id=41430 (дата обращения 17.09.2017)

5. T. Nakashika, T. Takiguchi, and Y. Ariki. Voice conversion using speaker-dependent conditional restricted boltzmann machine. EURASIP J. Audio, Speech and Music Processing, 2015:8, 2015.

6. Z. Wu. Spectral Mapping for Voice Conversion. PhD thesis, Nanyang Technological University, 3 2015.5.

7. Z. Wu, T. Virtanen, T. Kinnunen, E. Chng, and H. Li. Exemplar-based unit selection for voice conversion utilizing temporal information. In INTERSPEECH 2013, 14th Annual Conference of the International Speech Communication Association, Lyon, France, August 25-29, 2013, pages 3057–3061, 2013.