КОНВЕРСИЯ ГОЛОСА НА ОСНОВЕ ВЗВЕШЕННОЙ ДЕФОРМАЦИИ СПЕКТРА

Захарьев В. А., Петровский А. А.
Кафедра электронных вычислительных средств
Белорусский государственный университет информатики и радиоэлектороники
Минск, Республика Беларусь
E-mail: {zahariev, palex}@bsuir.by

В докладе предлагается методика построения системы конверсии голоса на базе модифицированной функции преобразования с использованием взвешенной деформации спектра, которая объединяет достоинства классических методов на основе моделей множественных гауссовых смесей и деформации спектра. Эксперименты показали эффективность данной методики, которая позволяет достичь высоких оценок степени соответствия исходного и целевого дикторов, натуральности и качества восстанавливаемого речевого сигнала, превосходящим качество классических методов на основе гауссовых смесей.

Введение

Конверсия голоса – это технология, которая используется для преобразования параметров речевого сигнала, характеризующих голос исходного диктора (ИД), таким образом, чтобы он был воспринят слушателями, как произнесённый другим человеком, целевым диктором (ЦД). Среди всех дикторозависимых характеристик голоса, конверсия голоса нацелена на изменение акустических характеристик: спектральных характеристик речи и параметров частоты основного тона [1]. В процессе обучения системы, на основе некоторого количества тренировочных данных полученных от конкретной пары ИД и ЦД, система определяет параметры функции конверсии на основе которой осуществляется оптимальное преобразование одного голоса в другой. Тип функции конверсии должен быть определен разработчиком системы заранее. В дальнейшем, непосредственно в процессе работы системы данная функция может быть применена для трансформации новых фраз поступающих на вход системы от ИД, для трансформации параметров сигнала в параметры ЦД [2–3].

Технология конверсии голоса имеет широкий спектр применения, в том числе: разработка мультиголосовых систем синтеза речи по тексту, персонализации звуковоспроизводящих устройств и программ, медицинские приложения для помощи людям с нарушениями речевого аппарата, дублирование кинофильмов, создание виртуальных клонированных голосов для видеоигр, мультимедиа и др.

I. Конверсия на основе МГС

Ядром системы конверсии голоса, является функция конверсии, на основе которой осуществляется непосредственное преобразование вектора параметров характеризующих голос ИД, в параметры ЦД. От её вида зависит степень приближения, т.е. качество работы системы конверсии. Вид функции (линейная, нелинейная и др.) и её параметры, в качестве которых могут выступать различные величины, в зависимости выбранной модели представления пространства акустических параметров диктора, должны быть определены разработчиком заранее.

На данный момент самой распространённой моделью является статистическая модель на основе множественных гауссовых смесей (МГС), которая в оригинале была предложена [4], а её доработанные варианты представлены [5]. Системы на базе данной модели имеют удовлетворительные результаты в плане характеристики сходства между преобразованным и целевым речевыми сигналами.

Обучение производится на основе набора параллельных пар векторов параметров ИД и ЦД,для которых строиться совместная модель МГС. Веткора математических ожиданий и ковариационные матрицы предоставляемы МГС используются в качестве параметров функции конверсии, которая представлена выражением:

$$F(\mathbf{x}) = \sum_{i=1}^{M} p_i(\mathbf{x})[\mu_i^y + \Sigma_i^{yx}(\Sigma_i^{xx})^{-1}(\mathbf{x} - \mu_i^x)], \quad (1)$$

$$p_i(\mathbf{x}) = \frac{\alpha_i N(\mathbf{x}, \mu_i^x, \Sigma_i^{xx})}{\sum_{j=1}^{M} \alpha_j N(\mathbf{x}, \mu_j^x, \Sigma_j^{xx})}.$$

где $\mathbf{x}$ – вектор параметреров исходного диктора,M – количество компонент смеси, μ_i^x и μ_i^y – вектора мат. ожиданий i-ой компоненты смеси, Σ_i^{xx} – ковариационная матрица исходного диктора i-ой компоненты, Σ_i^{yx} – кроссковариационная матрица для векторов исходного и целевого диктора i-ой компоненты, $p_i(\mathbf{x})$ –апостореорная вероятность принадлежности вектора $\mathbf{x}$ i-ой компоненте, N – многомерное Гауссово распределение с параметрами преведёнными выше.

Функция конверсии данного вида получила широкое распространение, благодаря хорошим практическим результатам полученным на её основе [5]. Однако, она также не лишена некоторых недостатков. При попытке увеличения степени сходства наблюдается эффект деградации

восстановленного речевого сигнала. Анализ различных моделей показывает, что одной из основных трудностей при разработке новых методов КГ является необходимость соблюдения данного компромисса между показателями сходства голоса и качества сигнала. Для решения данной проблемы была предложена модифицированная функция конверсии на основе МГС и взвешенной динамической трансормации спектра, речи о которой и поёдёт в данном докладе.

II. Конверсия на основе ВДС

В процессе экспериментов было установлена высокая степень корреляции между усреднёнными спектральными огибающими, характеризующимися векторами средних значений параметров μ_i^x и μ_i^y ИД и ЦД для каждой из компонент МГС. Поэтому было выдвинуто предложено, что возможно применение техники взвешенно деформации спектра (ВДС) в виде специальной функции конверсии, вместо простой линейной функции отображения параметров, на основе которой построено выражение (1). Это наблюдение вызвало интерес, поскольку использование частотной деформации должно обеспечить хорошую степень сходства без значительной деградации восстанавливаемого сигнала, что следует из определенных свойств данного преобразования [7]. Поскольку все вектора параметров внутри каждой компоненты МГС содержат параметрическое представление фонем со сходной формантной структурой, было предположено что единственная функция деформации частоты может быть использована для всех векторов принадлежащих данному классу модели.

В результате метод конверсии голоса, на основе взвешенно деформации спектра может быть описан следующим образом. Во время обучения, после того как были оценены характеристики совместной МГС на основе параллельной последовательности векторов параметров, усреднённые спектры определяемые μ_i^x и μ_i^y используются для построения функции деформации $W_i(f)$ для каждой компоненты смеси (см. рис. 1).

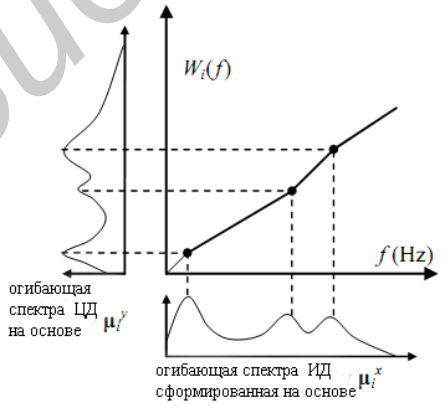

Рис. 1 – Функция деформации для усреднённого спектра i-й компоненты Гауссовой смеси

В процессе конверсии, на основе спектра k-го фрейма $S^{(k)}(f)$ для ИД и его параметрического представления $\mathbf{x^{(k)}}$ [6], применяется следующее преобразования для получения результирующего сконвертированного спектра $S'^{(k)}(f)$:

$$S'^{(k)}(f) = G^{(k)}(f) \cdot S_{fw}^{(k)}(f)$$

где $S_{fw}^{(k)}(f)$ – частотно-модифицированная версия оригинального спектра, рассчитываемая с применением взвешенной комбинации траекторий деформации $W_i(f)$,

$$S_{fw}^{(k)}(f) = S^{(k)}(W^{(k)^{-1}}(f))$$

$$W^{(k)}(f) = \sum_{i=1}^{M} p_i(\mathbf{x^{(k)}}) \cdot W_i(f)$$

и $G^{(k)}(f)$ корректирующий сглаживающий фильтр, который компенсирует разницу амплитуд деформированного спектра $S_{fw}^{(k)}(f)$ и реального целевого спектра, вычисляемого с использованием вектора $F(\mathbf{x}^{(k)})$ согласно выражению (1).

Заключение

Проведенные эксперименты по определению объективных и субъективны оценок качества показали, что использование взвешенной деформации спектра позволяет значительно улучшить качество системы, по сравнению со стандартными методами на базе МГС. В среднем по критерию соответствия дикторов оценку удалось поднять на 0,7 единицы в диапазоне от 1 до 5 по шкале MOS.A средний уровень качества восстанавливаемого сигнала составил 3,5 единицы, что доказывает возможность успешного применения данного метода в реальных приложения.

1. Shikano, K. Speaker adaptation through vector quantization / K. Shikano, K. Lee, R. Reddy // ICASSP. – 1986. –Vol. 11. –P. 231–237.
2. Stylinau, Y. Voice tracnsformation: A survey / Y. Stylinau. // ICASSP. –2009. –P. 3585–3588.
3. Voice conversion: a crtitical survey [Electronic resource] /A. F. Machado, M. Queiroz // Open-acess article. –2010. –modeof acces: http://www.ime.usp.br/ mqz/SMC2010Voice.–Date of access: 13.05.2013.
4. Toda, T. Spectral conversion based maximum liklehood estimation considering global variance of converted parameter / T. Toda, A. Black, K. Tokuda. // ICASSP. –2005. –P. 9–12.
5. Stylinau, Y. Continuous probabilistic transform for voice conversion. / Y. Stylianou // IEEE TSAP. – 1998. –№ 6 –P. 131–142.
6. Анализаторы речевых и звуковых сигналов /под ред. д.т.н. профессора Петровского А.А. –Минск: Бестпринт, 2009. –456 с.
7. Erro, D. Weighted frequency warping for voice conversion / D. Erro, A. Moreno // Audio, Speech, and Language Processing, IEEE Transactions on – 2010. – Vol. 18. –P. 543–550.