УДК 004.934

КОНВЕРСИЯ ГОЛОСА НА ОСНОВЕ МНОЖЕСТВЕННОЙ РЕГРЕССИОННОЙ ФУНКЦИИ ОТОБРАЖЕНИЯ

В.А. ЗАХАРЬЕВ, А.А. ПЕТРОВСКИЙ

Белорусский государственный университет информатики и радиоэлектроники
ул. П. Бровки, 6, г. Минск, 220013, Республика Беларусь
zahariev@bsuir.by, palex@bsuir.by

В докладе рассматриваются вопросы развития методов конверсии голоса. На основе статистических моделей, предлагается расширение стандартной функции конверсии путём выявления и учёта большего количества корреляционных связей между характеристиками дикторов и увеличению числа предикторов в составе множественной регрессионной функции отображения.

Ключевые слова: конверсия голоса, гауссовы смеси, множественная регрессионная модель.

Конверсия голоса (КГ) – это технология обработки речевого сигнала, позволяющая реализовать процесс трансформации параметров голоса, характеризующих речь исходного диктора (ИД), в параметры целевого (ЦД) [1]. Её центральной задачей является поиск функции конверсии, позволяющей выполнить оптимальное отображение вектора параметров, характеризующих речь исходного диктора, на каждом фрейме анализируемого сигнала, в параметры целевого диктора. Наборы таких векторов параметров для ИД и ЦД образуют соответствующие акустические пространства. В качестве критерия оптимальности преобразования, как правило, выступает минимум расстояния между сконвертированными векторами в данных пространствах. Функция КГ может быть реализована с использованием различных методов и моделей представления о том, какими характеристиками обладают и как между собой связаны оба пространства параметров ИД и ЦД [1].

Наиболее популярным из статистических методов, доказавших свою эффективность применения, является статистическая модель на основе множественных гауссовых смесей (МГС) [2]. Данная модель позволяет выполнить нежёсткую классификацию пространств дикторов, с учётом того что классы могут перекрываться, а также построить непрерывную функцию конверсии основанную на мягкой классификации [3]. Результаты конверсии на основе данных функций выгодно отличались от конверсии на основе подходов с жёсткой кластеризацией пространства параметров, например, векторного квантования, поскольку позволяли избежать возникновения артефактов в выходном речевом сигнале [4]. Функция конверсии на основе модели МГС может быть представлена:

$$F(x) = \sum_{q=1}^{Q} p_q(x)[\mathbf{v}_q + \mathbf{\Gamma}_q \mathbf{\Sigma}_q^{-1}(x - \boldsymbol{\mu}_q)],$$

где $p_q(x)$ – апостериорная вероятность того что входного вектор x принадлежит q-ой гауссовой компоненте, $\mathbf{\Gamma}_q$ – кроссковариационная матрица для векторов ИД и ЦД, а $\mathbf{v}_q$ – вектор средних значений для ЦД q-ой компоненты соответственно. Параметры $\{\mathbf{v}_q, \mathbf{\Gamma}_q\}$ вычисляются с применением методов среднеквадратической оптимизации с целью минимизации ошибки преобразования между сконвертированными и целевыми векторами на обучающей выборке. Эксперименты над системой построенной на базе

представленной выше функции показали хорошие результаты. Однако, является очевидным тот факт, что данный метод рассматривает последовательность векторов обучения как простой набор элементов для которых статистические связи присутствуют лишь для одной пары в каждый *i-ый* момент времени (рис. 1, *а*). Таким образом, данный метод учитывает исключительно пространственную корреляцию между векторами параметров без учета того, что параметры речевого сигнала не изменяются мгновенно, и поэтому обладают некоторой эргодичностью, что позволяет учесть не только пространственные но и континуальные корреляционные между смежными векторами параметров речевого сигнала (рис 1, *б*).

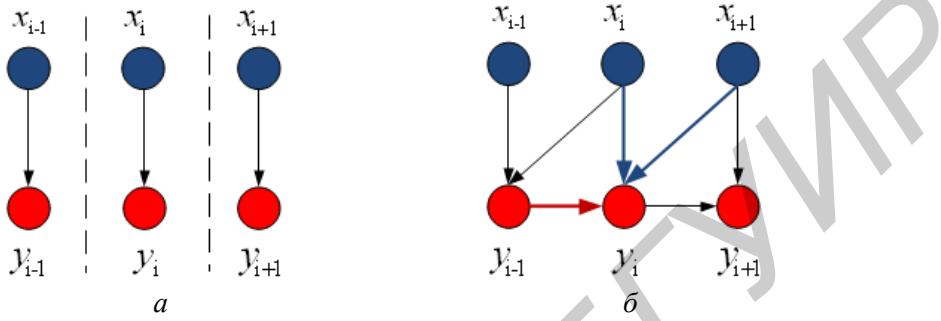

Рис. 1. Виды зависимостей между парами векторов обучающей последовательности:
а – независимая модель; *б* – Марковский процесс

В докладе предложена эргодическая моделью, которая учитывает зависимость в последовательности векторов не только для исходного, но и для целевого диктора, придавая тем самым последовательности свойства Марковского процесса. Таким образом, если элементы обучающей выборки целевого диктора условно считать состояниями модели, то регрессия учитывает следующие состояния

$$\mathbf{y}_i = \sum_{q=1}^{Q} p_q(\mathbf{x}_i, \mathbf{y}_{i-1}, \mathbf{x}_{i+1})\left[\mathbf{v}_q + \mathbf{\Phi}_q \bar{\mathbf{x}}_i^q + \mathbf{\Psi}_q \bar{\mathbf{y}}_{i-1}^q + \mathbf{\Omega}_q \bar{\mathbf{x}}_{i+1}^q\right],$$

Для определения параметров множественной регрессионной функции метод на основе совместной плотности вероятности применён быть не может. Поэтому поиск коэффициентов данной модели ведётся на основе более общего метода на базе наименьших квадратов. Таким образом, за счёт введения в регрессионную модель новых предикторов, учитывающих дополнительную статистическую информацию, и расширения функции конверсии, удаётся более эффективно осуществлять конверсию голоса, без внесения существенных артефактов в результирующий речевой сигнал, при этом сохранив высокие характеристики узнаваемости.

Список литературы

1. Moulines, E. Voice conversion: State of the art and perspectives / E. Moulines, Y. Sagisaka // Speech Communication. – 1995. – Vol. 16. – P. 125–224.

2. Bishop, C. M. Pattern Recognition and Machine Learning (Information Science and Statistics) / Christopher M. Bishop. – Secaucus, NJ, USA : Springer-Verlag New York, Inc., 2007. – 738 p.

3. Stylianou, Y. Statistical methods for voice quality transformation. / Yannis Stylianou, Olivier Cappe, Eric Moulines // Proc. of European Conference on Speech Communication and Technology. – Madrid, 1995. – P. 447–450.

4. Arslan, L. Speaker transformation algorithm using segmental codebooks (STASC) / L. Arslan // Speech Communication. – 1999. – Vol. 28, no. 3. – P. 211–226.