Скрытые марковские модели для решения задач текстонезависимого обучения в системах конверсии голоса

Захарьев В. А.
Кафедра электронных вычислительных средств
Белорусский государственный университет информатики и радиоэлектроники
Минск, Беларусь
e-mail: zahariev@bsuir.by

***Аннотация*—Конверсия голоса является одним из быстрорастущих и развивающихся направлений в области речевых технологий. В докладе рассмотрены вопросы создания систем конверсии с текстонезависимым обучением. В таких системах априорная речевая информация, поступающая от исходного и целевого дикторов и используемая для обучения системы, может существенно различаться в фонетическом и просодическом плане. Широкое распространение при решении данной задачи получили статистические модели, на основе марковских процессов. В частности, скрытые марковские модели. В докладе приведены основные понятия, определения и принципы построения и использования скрытых Марковских моделей. Рассматриваются конфигурации и параметры моделей оптимальные для решения задачи текстонезависимого обучения в системах конверсии голоса.**

Ключевые слова: конверсия голоса; текстонезависимое обучение; скрытая Марковская модель; параметрическое представление речевого сигнала.

I. Введение

Конверсия голоса – это процесс преобразования параметров речевого сигнала, характеризующих голос одного диктора, в параметры другого, без изменения лингвистической составляющей самого сообщения. Первый диктор называется исходным, второй – целевым. Процесс конверсии голоса подразумевает изменение акустических, фонетических, и просодических характеристик исходного диктора в характеристики целевого согласно определенному выражению (или набору правил), представляющему собой функцию конверсии [1]. Процесс функционирования системы конверсии включает в себя два основных этапа: обучения системы, и непосредственно, конверсии.

На первом этапе производится обработка априорной информации, в процессе которой происходит: параметризация входных сигналов исходного и целевого дикторов на основании выбранной модели речеобразования, их временное выравнивание друг относительно друга, формирование кластеров признаков описывающих фонетические и просодические свойства дикторов и построение на их основе функции конверсии.

На втором этапе также необходима параметризация речевого сигнала, однако, уже только исходного диктора. Вектор его параметров должен быть преобразован функцией конверсии, полученной на первом этапе. В заключении по нему должен быть синтезирован речевой сигнал с характеристиками голоса целевого диктора [2].

Из данных рассуждений легко видеть, что основной интерес для исследователей представляет именно этап обучения, в этой плоскости лежат основные задачи конверсии. От принципов построения данного этапа также зависит тип обучения – текстозависимый или текстонезависимай.

При текстозависимом обучении исходный и целевой дикторы обязаны начитывать один и тот же текст. Общий текст подразумевает общее фонетическое составляющее, варьирующееся в определенных пределах (в зависимости от пола, возраста, манеры и др. особенностей диктора), однако, принципиально одинаковое с точки зрения фонетики, что дает возможность провести временное сопоставление векторов признаков, а в дальнейшем сопоставить кластеры акустических подпространств дикторов.

В случае же текстонезависимого обучения дикторы вольны начитывать в микрофон произвольный текст (возможно, даже на разных языках). В результате чего явного временного соответствия векторов параметров прямым сопоставлением получить не возможно. Для решения данной проблемы прибегают к использованию статистических моделей, в том числе, скрытых Марковских моделей.

II. Скрытые марковские модели

Цепь маркова или марковский процесс – это последовательность событий, называемых состояниями, вероятности наступления которых зависят только от того в каком из состояний система пребывала до этого. Если через $S = \{s_1, s_2, ..., s_N\}$ обозначить вектор состояний в которых может находиться система, а через q_t состояние системы в момент времени $t = 1, 2, ...$ то вероятность нахождения системы $P(q_t \mid q_{t-1}, q_{t-2}, ...) = P(q_t \mid q_{t-1})$.

Скрытая марковская модель (СММ) является стохастической последовательностью, типа Марковской цепи, в которой состояния модели, на прямую не наблюдаются, а являются некоторой вероятностной функцией данного состояния. Иными словами СММ представляет собой дважды

стохастический процесс, один из которых, в виде переходов модели из одного состояния в другое $Q = \{q_1, q_2, ..., q_K\}$, является основным и ненаблюдаемым (т.е. скрытым). Единственное, что мы можем, – это судить о нем с помощью другого случайного процесса, который нам даёт последовательность наблюдений $X = \{x_1, x_2, ..., x_K\}$.

Для определения СММ необходимо задать следующие элементы: N – число состояний модели, M – число различных символов наблюдений (либо интервал значений наблюдаемого параметра для непрерывных моделей), распределение вероятностей переходов между состояниями (1), распределение вероятностей появления символов в j-ом состоянии (2), начальное распределение состояний (3):

$$A = \{a_{ij}\}^{N \times N},\ a_{ij} = P(q_{t+1} = S_j \mid q_t = S_i),\ 1 \le i, j \le N \quad (1)$$

$$B = \{b_j\}^{N \times M},\ b_j = P(v_k \mid q_t = S_j),\ 1 \le j \le N, 1 \le k \le M \quad (2)$$

$$\pi = \{\pi_i\}^{N \times 1},\ \pi_i = P(q_i = S_i), 1 \le i \le N. \quad (3)$$

После задания начальных значений, при условии наличия обучающих выборок наблюдений, может производиться уточнение параметров модели с использованием итерационных процедур Витерби и Баума – Велша [3].

III. СММ для решения задачи текстонезависимого обучения

Обобщенный процесс работы модуля текстонезависимого обучения представлен на Рис. 1. Текст, произносимый дикторами, состоит из последовательности фонем. В данном случае полагается их соответствие состояниям СММ, поэтому обозначим данные последовательности $S^{src} = \{s_1^{src}, s_2^{src}, ..., s_N^{src}\}$ – для исходного диктора, $S^{trg} = \{s_1^{trg}, s_2^{trg}, ..., s_N^{trg}\}$ – для целевого диктора. Каждая из фонем имеет непосредственную реализацию в речевом потоке, которая может быть различной в зависимости от большого кол-ва внешних факторов, и по-разному проявляется в параметрах речевого сигнала. После параметризации речевого сигнала, необходимо определить соответствие определенного набора векторов параметров, определенной последовательности фонем, а также сопоставить фонему s_i с последовательностью наблюдений $\{x_i\}$, – распознать состояния СММ [4].

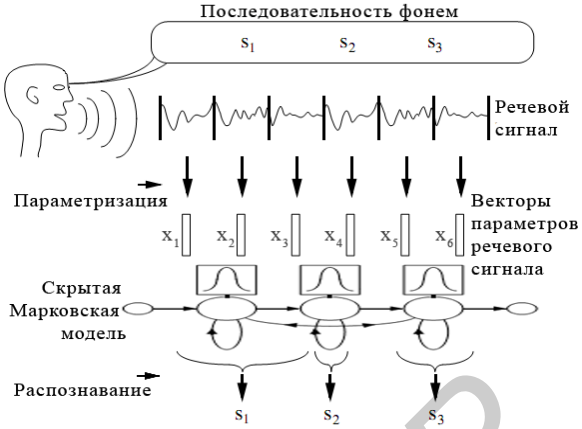

Рис. 1. Процесс текстонезависимого обучения системы конверсии с использованием СММ

С точки зрения СММ данный процесс заключается в том, чтобы связать оптимальную последовательность состояний с текущей последовательностью наблюдений для данной модели, что можно записать с помощью следующего математического выражения (4):

$$\begin{cases} \arg\max P(X^{src}, S^{src} \mid \Theta^{src}),\ \Theta^{src} \in (A^{src}, B^{src}, \pi^{src}) \\ \arg\max P(X^{trg}, S^{trg} \mid \Theta^{trg}),\ \Theta^{trg} \in (A^{trg}, B^{trg}, \pi^{trg}) \end{cases}, \quad (4)$$

После решения данной задачи с помощью итерационного алгоритма Витерби. Существует возможность найти соответствие последовательностей фонем исходного и целевого $S^{src} \sim S^{trg}$ дикторов, а также векторов параметров к ним относящихся. Это позволяет сформировать совместный вектор параметров, для его последующей кластеризации, определения параметров функции конверсии, и, как следствие, успешного решения задачи обучения системы конверсии.

IV. Выводы

В процессе обучения и тестирования системы использовалась речевая база, содержащая звукозаписи различных текстов русскоязычных дикторов с битрейтом 256 кбит/с. Использовалась фонетическая разметка для обучения СММ. При условии фонетической сбалансированности начитанных текстов. Оптимальное параметры СММ: количество состояний 42 (по кол-ву фонем русского языка), наблюдения моделировались непрерывной плотностью вероятности в виде моделей Гауссовых смесей с кол-во компонент 8, кол-во обучающих фраз составляли по 10 ш. на одного диктора.

[1] K. Shikano, K. Lee, and R. Reddy, “Speaker adaptation through vector quantization”, ICASSP, vol .11, 1986.

[2] Петровский А.А. “Анализаторы речевых и звуковых сигналов”, Минск: Бестпринт, - 2009, - 456 с.

[3] Рабинер Л.Р, “Скрытые марковские модели и их применение в избранных приложениях при распознавании речи: Обзор”, ТИИЭР, т. 77, № 2, февраль 1989, СС. 86-120.

[4] Bourlard H. “Introduction to Hidden Markov Models”, Ecole Polytechnique de Lauseane, - 2010, - 120 p.