УДК 519.688

ТЕХНОЛОГИЯ СИНТЕЗА РУССКОЙ РЕЧИ НА ОСНОВЕ СКРЫТЫХ МАРКОВСКИХ МОДЕЛЕЙ

П.Г. Чистиков

Представлен подход к построению системы синтеза речи на основе скрытых марковских моделей применительно к русскому языку. Для повышения ее гибкости применяется алгоритм кластеризации состояний. Представлен подход моделирования сигнала возбуждения.

Ключевые слова: синтез речи, скрытые марковские модели, параметризация речи, кластеризация данных.

Архитектурно и логически систему синтеза можно разделить на две части – построение модели голоса и непосредственно синтез [1]. Первая часть включает в себя следующие этапы: вычисление акустических, лингвистических и просодических признаков для каждого аллофона из базы данных; обучение контекстно-зависимых HMM (скрытых марковских моделей); кластеризация состояний HMM на основе лингвистических и просодических признаков. Синтезирующая часть представляет собой следующую последовательность действий: транскрибирование входного текста и вычисление лингвистических и просодических характеристик для каждого аллофона; формирование последовательности HMM; генерация акустических параметров на основе полученной модели; вычисление функции возбуждения и ее фильтрация с целью получения итогового речевого сигнала.

Неотъемлемой составляющей для построения системы синтеза речи является выбор параметров, обеспечивающих генерацию естественного сигнала. Такие параметры могут включать, например, тип предыдущего/следующего аллофона, слога, слова, предложения и т.д. Определение набора таких параметров для определенного языка базируется на лингвистической и просодической информации. Помимо теоретического подхода, применяется также эмпирический анализ с целью выбора наиболее информативных из них. Так, для русского языка было выбрано 7 аллофонных, 13 слоговых, 8 словных и 3 синтагматических признака, таких как имя фонемы, предшествующей предыдущей, имя предыдущей фонемы, имя текущей фонемы, имя следующей фонемы, имя фонемы, следующей за следующей, позиция текущей фонемы от начала слога, позиция текущей фонемы от конца слога и т.д.

Моделируемые параметры идеологически делятся на две группы – спектральные и временные. В качестве спектральных используются частота основного тона и мел-частотные кепстральные коэффициенты. Расчет данных параметров выполняется по всей фонограмме из базы данных с окном анализа 25 мс и смещением 10 мс. Временные параметры представляют собой длительности соответствующих аллофонов.

Важным этапом для обеспечения качественного синтезированного сигнала является моделирование функции возбуждения. Наиболее качественную работу показывает алгоритм [2], основанный на моделировании формы этой функции при помощи двух фильтров (единичных импульсов – (1) и белого шума – (2)),

$$H_v(z) = \sum_{l=-M/2}^{M/2} h(l) z^{-l}, \quad (1)$$

$$H_u(z) = \frac{1}{1 - \sum_{l=1}^{L} g(l) z^{-l}}, \quad (2)$$

для вокальной и шумовой составляющих соответственно, коэффициенты которых $h(l)$ и $g(l)$ вычисляются на этапе обучения. Порядки фильтров M и L равны 512 и 256 соответственно. Примеры синтеза показывают, не вдаваясь в детали качества воспроизведения аллофонов, что ритмика фразы сохраняется. Данный факт демонстрирует важную характеристику основанного на HMM синтеза речи: возможность имитировать просодические характеристики корпуса, который был использован при построении модели голоса. Также стоит отметить, что для построения модели голоса достаточно небольшого количества материала, однако отсутствие некоторых элементов в базе данных существенно влияет на качество, что делает процесс подготовки звуковой базы данных также очень важным при разработке систем синтеза.

Автором предложен подход к построению системы синтеза русской речи на основе скрытых марковских моделей. Принцип основан на методе, в котором соответствующие параметры извлекаются из скрытых марковских моделей, векторы наблюдений которых содержат спектральные характеристики, значения основного тона и длительности речи. Экспериментальные результаты показывают, что русская речь может быть успешно параметризована и произвольное предложение может быть синтезировано из полученных моделей.

1. Maia R., Zen H., Tokuda K., Kitamura T., Resende F.G. Towards the development of a Brazilian Portuguese text-to-speech system based on HMM // Proceedings of the European Conference on Speech Communication and Technology (EUROSPEECH). – Geneva, Switzerland, 2003. – P. 2465–2468.
2. Maia Ranniery, Toda Tomoki, Zen Heiga, Nankaku Yoshihiko, Tokuda Keiichi. An Excitation Model for HMM-Based Speech Synthesis Based on Residual Modeling // 6th ISCA Workshop on Speech Synthesis. – Bonn, Germany, 2007. – P. 1315–1318.

Чистиков Павел Геннадьевич – ООО «ЦРТ», научный сотрудник, аспирант, chistikov@speechpro.com