Анализ подходов конверсии голоса в системах мультимедиа

Захарьев В.А.
Кафедра ЭВС, ФКП
БГУИР
Минск, Беларусь
e-mail: zahariev-vadim@yandex.ru

***Аннотация* — В научных публикациях посвященных сопоставлению моделей конверсии речи, сравнительный анализ, как правило, проводится на основе стандартного ряда объективных и субъективных тестов. В данных работах отсутствуют признаки позволяющие сделать выводы о трудоёмкости алгоритмов, количественному и качественному составу данных необходимых для обучения моделей конверсии, что затрудняет выбор модели для приложений, где данные признаки имеют критическое значение. В статье представлен анализ эффективности использования моделей не только по результатам их работы, но и по составу априорных данных необходимых для их обучения. Такой анализ актуален при построении мультимедиа систем реального времени, а также систем с дружественным человеко-машинным интерфейсом и высоким уровнем юзабилити.**

Ключевые слова: конверсия речи; модель конверсии; системы мультимедиа; системы реального времени

I. Введение

Конверсия речи – это процесс преобразования параметров речевого сигнала одного диктора (исходного) в параметры другого диктора (целевого), без изменения лингвистической (информационной) составляющей самого сообщения. Данный процесс подразумевает изменение акустических, фонетических, и просодических характеристик исходного диктора в характеристики целевого согласно определенному набору правил, представляющему собой модель конверсии речи. Перцептивное качество преобразования определяется точность и сложностью построения модели конверсии, а также тем насколько хорошо могут быть аппроксимированы параметры целевого диктора параметрами исходного. Построение модели сводится к двум основным этапам.

На первом этапе производится обработка априорной информации, в процессе которой из исходных речевых сигналов обоих дикторов с использованием методов анализа (линейное предсказание, формантный или кепстральный анализ) речи происходит выделение характеристических векторов признаков достаточных для идентификации обоих дикторов. Затем, как правило, их временное масштабирование и выравнивание друг относительно друга.

На втором этапе происходит кластеризация пространства векторов признаков с использованием методов машинного обучения (векторное квантование, модели гауссовых смесей). Далее формируется правило конверсии – сердце всей системы – по которому, вектор признаков пространства исходного диктора преобразуется в наиболее соответствующий ему вектор пространства целевого диктора.

В качестве обучающих выборок используется отрезки речевого сигнала с одинаковой лингвистической составляющей, поэтому такие системы называются текстозависимыми системами конверсии.

В дальнейшем, в процессе работы системы, подразумевается, что обученная модель конверсии может преобразовывать голос исходного диктора в целевой на любом отрезке произвольной речи, в не зависимости от содержания.

Существует большое количество различных моделей конверсии основанных на методах цифровой обработки сигналов (векторное квантование [1], частотное масштабирование), линейной алгебры (линейная регрессия, сингулярное разложение векторов), статистических (скрытые Марковские модели [10], модели гауссовых смесей [6]), и когнитивных методологиях (искусственные нейронные сети [4], байесовский подход).

Однако, великое множество их реализаций и сочетаний, вкупе с использованием различных методов анализа речи, дающих в каждом конкретном случае различные результаты, а также отсутствие единых методик тестирования, разработанных непосредственно для систем конверсии голоса, затрудняет их сравнение, опираясь только лишь на результаты, полученные в ходе стандартных тестов для систем распознавания и синтеза речи. Поэтому в работе предлагается при сравнении качества моделей систем конверсии учитывать не только выходные характеристики систем, но и состав данных, сложность структуры и алгоритмов самих моделей.

II. Модели конверсии голоса

Среди большого количества моделей представленных выше, для проведения сравнения было выбрано три наиболее распространенных. Это векторное квантование, гауссовы смеси и нейронные сети.

A. *Векторное квантование*

Процедура векторного квантования подробно изложена в [2] или [5]. Здесь приведем лишь функцию конверсии которая имеет вид (1)

$$F(\vec{y}_t) = \sum_{i=1}^{N} p_i \vec{c}_i \,, \tag{1}$$

$$p_i = \frac{e^{-d_i}}{\sum_{j=1}^{N} e^{-d_j}}, \tag{2}$$

$$d_i = \sum_{k=1}^{N} v_k \left| \vec{c}_i - \vec{y}_t \right|, \tag{3}$$

где p_i – вес, характеризующий вероятность принадлежности вектора целевого диктора y_t к акустическому классу, представленному в кодовой книге исходного диктора размерностью N центроидой c_i, рассчитывается по формуле (2). А d_i является мерой искажения, рассчитываемой по формуле (3).

B. *Модель гауссовых смесей*

Впервые гауссовы смеси были предложены для использования в системах конверсии голоса в [7]. Функция конверсии определяется как (4)

$$\hat{y}_i = E[\vec{y}_i \mid \vec{x}_i] = \sum_{i=1}^{M} h_i(\vec{x})[\mu_i^{\vec{y}} + \Sigma_i^{\vec{y}\vec{x}}(\Sigma_i^{\vec{x}\vec{x}})^{-1}(\vec{x} - \mu_i^{\vec{x}})] \quad (4)$$

$$h_i(\vec{x}) = \frac{\alpha_i N(\vec{x}, \mu_i^{\vec{x}}, \Sigma_i^{\vec{x}\vec{x}})}{\sum_{j=1}^{M} \alpha_i N(\vec{x}, \mu_j^{\vec{x}}, \Sigma_j^{\vec{x}\vec{x}})}, \quad (5)$$

где $\vec{x}_i$ и $\vec{y}_i$ – вектора параметров *i-ого* класса исходного и целевого дикторов соответственно, $\mu_i^{\vec{x}}$, $\mu_i^{\vec{y}}$ – средние вектора *i-ого* класса, $\Sigma_i^{\vec{x}\vec{x}}$ – ковариационная матрица *i-ого* класса исходного диктора, $\Sigma_i^{\vec{y}\vec{x}}$ – кроссковариационная матрица *i-го* класса для исходного и целевого дикторов, $h_i(\vec{x})$ – апостериорная вероятность принадлежности вектора $\vec{x}$ *i-му* классу, $N(\vec{x}, \mu_i^{\vec{x}}, \Sigma_i^{\vec{x}\vec{x}})$ – гауссово распределение со средним вектором $\mu_i^{\vec{x}}$ и ковариационной матрицей $\Sigma_i^{\vec{x}\vec{x}}$, α_i – весовой коэффициент, *M* – количество компонент гауссовой смеси.

C. *Искусственные нейронные сети*

Способы применения нейросетевых моделей для решения проблем конверсии голоса были предложены [4]. В работе [8]соответствующая тематика была развита. В нашем случае для анализа и сопоставления моделей конверсии мы использовали нейросетевую структуру представленную на рис. 1.

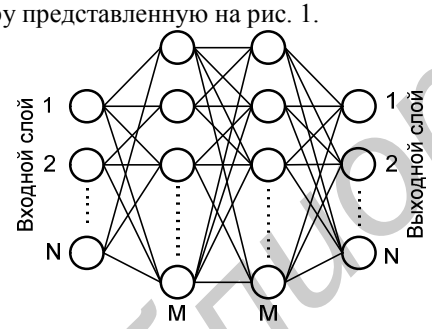

Рис. 1. Архитектура нейронной сети

Данная сеть имеет конфигурацию 25N-50M-50M-25N, т.е. имеет по 25 нейронов во входном и выходном слоях и по 50 нейронов в двух скрытых слоях. Используются линейные функции активации для входа и выхода, а также тангенциальные функции активации в скрытых слоях сети. Для обучения сети используется алгоритм обратного распространения.

III. Результаты анализа моделей

В ходе анализа первоисточников моделей, сравнения результатов тестов (*Mean Opinion Scores - MOS*) и (*ABX*) представленных в работах [3],[9],[11],[12],[13] собственных экспериментов, был определен минимальный объем словаря необходимого для обучения, а также выявлены следующие основные достоинства и недостатки моделей. Результаты представлены в Табл. 1.

Табл. 1. Резальтаты анализа моделей

Модель	***Объем словаря, ФСВ*** [a.]	***Достоинства***	***Недостатки***
Векторное квантование	100-200	– невысокая вычисл-ная сложность; – высокая степень подобия на ограниченных участках речи	– прерывная функция конвер.; –большой объем обуч данных; – большой объем памяти
Гауссовы смеси	30-50	– непрерывная функция конверсии; – возможность обучения по невыровненным векторам	– сложность определения кол-ва компонент; – необх. учета корреляц. связей между фреймами
Нейронные сети	30-40	– непрерывная функция отображен; – функция отображения действует на перех-ных сегмент.	– сложность выбора подходящ. архитектуры; – сложность выбора вида функции активации

a. ФСВ – фонетически сбалансированных высказываний

[1] D. G. Childers, B. Yegnanarayana, K. Wu, "Voice conversion:Factors responsible for quality", ICASSP,1985,pp.748-751,

[2] K. Shikano, K. Lee, and R. Reddy, "Speaker adaptation through vector quantization", ICASSP, vol .11, 1986.

[3] M. Abe, K. Shikano, H. Kuwabara, "Cross-language voice conversion", ICASSP, 1990, pp. 345-348.

[4] M. Narendranath, H. Murty, S. Rajendran, B. Yegnanarayana, "Transformation of formants for voice conversion using artificial neural networks", Speech Communication, vol. 16, 1995, pp. 207-216.

[5] L. Arslan, "Speaker transformation algorithm using segmental codebooks", Speech Communication, vol. 28, no. 3, 1999, pp. 211-226.

[6] Y. Stylianou, "Continuous probabilistic transform for voice conversion", IEEE TSAP, no. 6, 1998, pp. 131-142.

[7] T. Toda, H. Saruwatari, K. Shikano, "Voice conversion algorithm based on Gaussian mixture model with dynamic frequency warping of straight spectrum", Power [dB], vol. 30, 2001, pp. 40-48.

[8] S. Desai, E. V. Raghavendra, B.Yagnanarayana, A. W. Black, K. Prahallad, "Voice conversion using artificial neural networks"

[9] .Rentzos, S. Vaseghai, E. Turajlic, Q. Yan, C. Ho, "Transformation of speaker characteristics for voice conversion", IEEE WASRU, 2003, pp. 706-711.

[10] E. Kim, S. Lee, Y. Oh, "Hidden markov model based voice conversion using dynamic chracteristics of speaker", 5th ECSCT, 1997.

[11] T.Toda, A. Black, K. Tokuda, "Spectral conversion based maximum liklehood estimation considering global variance of converted parameter", ICASSP, 2005, pp 9-12.

[12] Y. Stylinau, "Voice tracnsformation: A survey", ICASSP, 2009, pp. 3585-3588.

[13] A. F. Machado, M. Queiroz, "Voice conversion: a crtitical survey", open-acces article, acces mode: www.ime.usp.br/~mqz/SMC2010_Voice.pdf, 2010, 8 p.