УДК 004.934

ПРЕОБРАЗОВАНИЕ ГОЛОСА НА ОСНОВЕ ГИБРИДНОГО SVR И GMM

Романов А. К.[1], Белов Ю.С.[1]

[1]Калужский филиал ФГБОУ ВО «Московский государственный технический университет им. Н.Э. Баумана (национальный исследовательский университет)», Калуга, e-mail: fn1-kf@mail.ru;

В данной статье представлен новый метод преобразования голоса, который основан на гибридном SVR (поддержка векторной регрессии) и GMM (модель смеси Гаусса). Преобразование голоса заключается в преобразовании характеристик голоса источника в характеристики голоса-цели. Доступны самые разнообразные технологии: от экспрессивного синтеза текста до речи с сохранением индивидуальности голоса в системе с низким уровнем шума. В течение последних десятилетий было предложено несколько методов преобразования голоса, таких как кодовая книга сопоставления, функция дискретной трансформации, GMM и ANN (искусственная нейронная сеть). Возможности отображения SVR и GMM используются для сопоставления спектральных характеристик исходного голоса с целевыми. Также представлена новая стратегия преобразования F0: модели F0 моделируются со спектральными характеристиками совместно с GMM и прогнозируются по преобразованным спектральным признакам с использованием метода SVR. Субъективные и объективные тесты проводятся для оценки работы. Экспериментальные результаты показывают, что с использованием данного метода преобразованная речь получается лучшего качества, чем при использовании современного метода GMM. Между тем, также предлагается метод VC, основанный на непараллельных данных, информация об источнике исследуется с использованием метода SVR.

Ключевые слова: преобразование голоса, SVR, GMM, преобразование F0, характеристика динамической информации.

VOICE CPNVERSION BASED ON HYBRID SVR AND GMM

Romanov A.K, Belov Yu.S.

[1]Moscow State Technical University n.a. Bauman (National Research University), Kaluga Branch, Kaluga, e-mail: fn1-kf@mail.ru;

A novel VC (voice conversion) method based on hybrid SVR (support vector regression) and GMM (Gaussian mixture model) is presented in the paper, the mapping abilities of SVR and GMM are exploited to map the spectral features of the source speaker to those of target ones. VC (voice conversion) is a technique which refers to transforming the characteristics of a source speaker to those of a target speaker. A wide variety of applications are available, ranging from expressive text-tospeech synthesis and preserving speaker individuality in an ultra low bit communication system, to aiding the speech-impaired people. Several VC methods have been proposed over the past decades, such as the mapping codebook, the discrete transformation function,GMM, and the ANN (artificial neural network). A new strategy of F0 transformation is also presented, the F0s are modeled with spectral features in a joint GMM and predicted from the converted spectral features using the SVR method. Subjective and objective tests are carried out to evaluate the VC performance; experimental results show that the converted speech using the proposed method can obtain a better quality than that using the state-of-the-art GMM method. Meanwhile, a VC method based on non-parallel data is also proposed, the speaker-specific information is investigated using the SVR method.

Keywords: voice conversion, SVR, GMM, F0 transformation, characteristic of dynamic information.

Введение

В течение последних десятилетий было предложено несколько методов преобразования голоса, таких как кодовая книга, дискретная функция преобразования, GMM и ANN (искусственная нейронная сеть). В методе сопоставления кодовой книги применяется технология кластеризации спектральных параметров исходного и целевого голоса, а функция отображения получается из двух результирующих кодовых книг. Одним из основных недостатков данного метода является то, что преобразованные

параметры имеют ограниченное число, что приводит к серьезному ухудшению качества речи. Далее, чтобы заменить метод кодовой книги сопоставления было предложено дискретное преобразование с использованием кусочно-линейной функции. Однако, использование этой технологии приводило к разрывам в преобразованной речи. В методе VC на основе GMM преобразование устанавливается на основе непрерывных вероятностных функций. Экспериментальные результаты показывают, что полученные результаты намного лучше, по сравнению с другими предыдущими методами трансформации. Метод ANN как непрерывной и нелинейной функции также исследован, и было доказано, что результаты, сопоставимые с результатами метода GMM, могут быть достигнуты, но есть несколько основных недостатков, таких как большая вычислительная нагрузка, множественные локальные минимумы в зависимости от минимизации эмпирического риска и связанной с проблемой преодоления. В настоящий момент GMM является самым популярным и хорошо зарекомендовавшим себя методом преобразования голоса. Было предложено много улучшенных методов на основе GMM, таких как метод GMM и DFW (метод динамической деформирования частоты), метод GMM и MAP (максимальный апостериорный метод) и GMM с использованием метода генерации параметров ML (максимального правдоподобия). Эти методы избегают проблем разрыва преобразованной речи. В отличие от традиционных методов GMM или ANN, SVR-подход может отображать нелинейную зависимость между исходным и целевым голосами, ему требуется меньше данных для обучения, и он менее подвержен локальным минимумам. Таким образом, предлагается гибридный метод преобразования голоса на основе SVR и GMM, в котором SVR-сопоставление выполняется вместо линейной регрессии в каждом компоненте GMM.

Базовый GMM на основе спектрального преобразования

Существует два основных подхода преобразования голоса, основанных на GMM, первый - LSE (метод наименьшего квадрата) и JDE (совместная оценка плотности). Эти методы показывают эквивалентную производительность, которая выбирается в качестве базовой линии метода преобразования голоса на основе гибридного SVR и GMM.

Пусть $x = \{x_1 \dots x_N\}$ и $y = \{y_1 \dots y_N\}$ - последовательности спектральных параметров исходного и целевого голоса соответственно, где $x_i = \{x_{i1} \dots x_{ij}\}$ и $y_i = \{y_{i1} \dots y_{ij}\}$ [1]. Последовательность x выровнена относительно y, чтобы получить пару параллельных последовательностей $z = (x^T, y^T)^T$ (где надстрочный индекс T обозначает транспонирование), которая используется для обучения совместных параметров GMM (α, μ, Σ). GMM можно записать в виде суммы M гауссовых компонент, которая принимает вид:

$$p(z) = \sum_{i=1}^{M} \alpha_i N(z,\ \mu_i,\ \sum_i) \tag{1}$$

где α_i обозначает предыдущую вероятность i-го компонента и удовлетворяет $\sum_{i=1}^{M} \alpha_i = 1$, μ_i и $\sum_i$ - средние и ковариационные матрицы i-го компонента [2]. Функция преобразования и минимизации средних квадратичных ошибок между преобразованной и целевой речью, может быть записана как:

$$F(x) = E(y|x) = \sum_{i=1}^{M} p_i(x) \left[\mu_i^y + \frac{\sum_i^{yx}}{\sum_i^{xx}} (x - \mu_i^x) \right] \tag{2}$$

$$p_i(x) = \frac{\alpha_i N(z,\ \mu_i^x,\ \sum_i^{xx})}{\sum_{k=1}^{M} \alpha_k N(z,\ \mu_k^x,\ \sum_k^{xx})} \tag{3}$$

где, $\mu = \begin{bmatrix} \mu_i^x \\ \mu_i^y \end{bmatrix}$ и $\sum_i = \begin{bmatrix} \sum_i^{xx} & \sum_i^{xy} \\ \sum_i^{yx} & \sum_i^{yy} \end{bmatrix}$ и pi (x) - вероятность x, принадлежащая i-му компоненту.

Предлагаемые гибридные методы SVR и GMM

Рассмотрим новый гибридный метод преобразования голоса на основе SVR и GMM. SVR применяется в каждом компоненте GMM, он радикально отличается от традиционных методов GMM или ANN. Данный метод выполняет идеальное нелинейное отображение между исходным и целевым голосами и может эффективно избегать проблемы с перегрузкой и всегда находит глобальные минимумы [3]. В отличии от традиционного одномерного выходного значения, SVR предлагает многомерное преобразование голоса [4]. Функция преобразования представляет собой регрессию в i-м компоненте и задается выражением:

$$f_m(x) = \left[W,\ \varphi(x)\right] + b \tag{4}$$

где $\varphi(x)$ - нелинейная функция отображения из низкоразмерного пространства в более высокое, $W = \{\omega_1 \dots \omega_J\}^T$ и $b = \{b_1 \dots b_J\}^T$ определяют два J-мерных регрессора в верхнем размерном пространстве. Функция регрессии может быть получена путем оптимизации:

$$\begin{gathered} min \frac{1}{2} \sum_{j=1}^{J} \left\| w_j \right\|^2 + C \sum_{i=1}^{N} L(\xi_i) \\ s.t.\ \left\| y_j - \left(W, \varphi(x_i)\right) - b \right\| \le \varepsilon + \xi_i \\ \xi_i \ge 0, i = 1, \dots N \end{gathered} \tag{5}$$

Здесь C - штрафной коэффициент, ε и ξ_i - переменные, которые показывают ошибками на обучающих точках, а L(ξ_i) обозначает функцию стоимости [5]. Вместо гиперкубической интенсивной зоны, используемой в SVR на основе ε, принимается гиперсферическая нечувствительная зона, чтобы справиться с многомерным выходом, а также метод IRWLS (итеративный реверсифицированный наименьший квадрат) используется для решения уравнения Лагранжа следующим образом:

$$L(W,b) = \frac{1}{2}\sum_{j=1}^{J}\|w_j\|^2 + C\sum_{i=1}^{N}L(\xi_i) - \\ -\sum_{i=1}^{N}\alpha_i\left[(\varepsilon+\xi_i)^2 - \|y_i - (W,\varphi(x_i)) - b\|^2\right] - \sum_{i=1}^{N}\mu_i\xi_i \quad (6)$$

где α_i и μ_i - множители Лагранжа. Неизвестные параметры w_j и b_j будут вычисляться в каждом измерении. Таким образом, функция преобразования голоса на основе GMM может быть модифицирована как:

$$F(x) = \sum_{i=1}^{M} p_i(x) f_i(x) \quad (7)$$

Как известно, правильный выбор ядра является ключом к производительности SVR. RBF (радиальная базисная функция) и полиномиальная функция являются двумя типичными методами ядра. RBF обладает лучшей интерполяционной характеристикой, в то время как полиномиальная функция K_p показывает лучшую экстраполяционную способность. Для повышения эффективности преобразования вводится смешанное ядро:

$$K_{mix} = \lambda K_{rbf} + (1-\lambda)K_p, \ \ 0 \le \lambda \le 1 \quad (8)$$

Вес λ изменяется от 0 до 1 при размере шага 0,05.

Преобразование F0

Типичный метод преобразования F0 основан на GMM, который описывается формулой 2. Многие исследования показали взаимосвязь между спектральными параметрами и F0. Совместное GMM используется для моделирования F0 и спектральных характеристик [6]. Предсказание F0 из MFCC (частотного векторного кепстрального коэффициента), где используются методы GMM и HMM (скрытая марковская модель) указывает на то, что эти методы могут достичь удовлетворительных результатов, как прогнозировалось, но все же существуют некоторые недостатки: линейность отношений между спектральными параметрами и F0, а также потребность в большом объеме данных

на этапе обучения [7]. Таким образом применяется метод SVR, который делает нелинейное сопоставление, используя меньше данных для обучения, вследствие чего F0 прогнозируется по спектральным параметрам. В зависимости от традиционных методов преобразования F0, для процесса обучения необходимы только целевые функции. Модификация F0 выполняется следующим образом:

Шаг 1. На этапе тренировки, последовательности спектральных параметров $y = \{y_1 \dots y_N\}$ и $f = \{f_1 \dots f_N\}$ целевой речи вычисляются с использованием метода STRAIGHT.

Шаг 2. Затем, алгоритм EM (ожидание максимизации) используется для моделирования y и f в совместном GMM. В каждом компоненте GMM функция преобразования обучается между y и f, используя SVR на основе ε со смешанным ядром.

Шаг 3. На этапе преобразования, F0 оценивается по преобразованным спектральным параметрам с использованием обучаемых функций преобразования SVR.

Преобразование голоса на основе непараллельных данных

Ранее обсуждавшиеся методы преобразования голоса были основаны, главным образом, на параллельных данных для обучения, которые требуют одинаковые исходные и целевые высказывания. Недавно были исследованы подходы с использованием непараллельных данных [8], они могут получить удовлетворительные результаты, но все еще нуждаются в некоторой предварительной информации от функции отображения между исходными и целевыми голосами, что не всегда возможно в реальных приложениях. В данной статье применялся метод SVR для записи специфичной голосовой информации, который не нуждается в какой-либо предварительной голосовой информации источника и позволяет делать преобразование голоса от произвольного источника до цели. Идея стимулируется специфическим отображением для распознавания речи [9]. Пусть L обозначает лингвистическую информацию, а LS соответствует лингвистической и говорящей информации. Функция отображения Q(L) вычисляется для получения отношений между L и LS и вычисляется с использованием метода LSE на данных обучения, чтобы минимизировать квадратичные ошибки:

$$SE = \sum_{i=1}^{N} ||LS_i - Q(L_i)||^2 \quad 9)$$

Предполагая, что m и n являются порядками для L и LS, которые трудно определить. Низкий порядок анализа LP (линейный предсказательный) (m: 4~8) может грубо фиксировать лингвистическую информацию говорящего, тогда как более высокий порядок LP (n:> 12) может захватывать как лингвистическую, так и динамическую информацию [10]. На рисунке 1 показана блок-схема процесса обучения SVR на основе

голосового преобразования с использованием специфичной информации, и метод VTLN (нормализация длины голосового тракта), поскольку модуль предварительной обработки принимается для извлечения лингвистической информации.

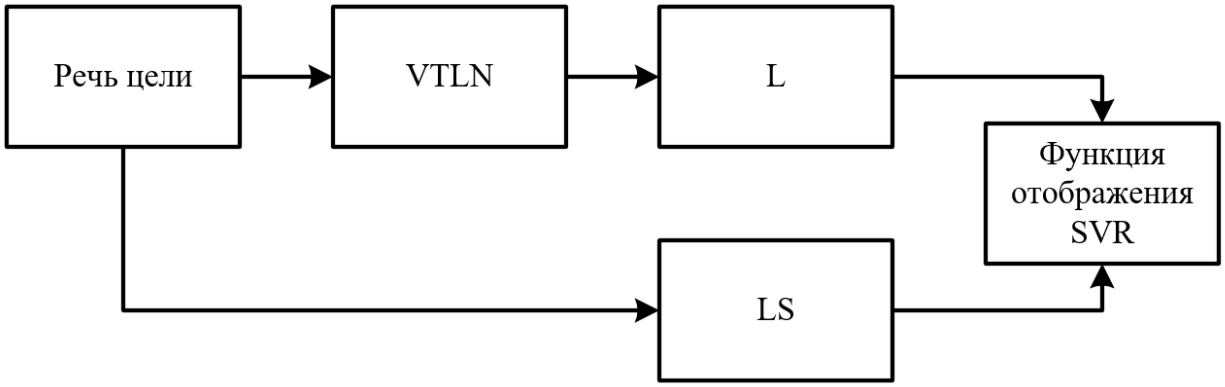

Рис. 1. Структура обучения системы преобразования голоса, использующая информацию о голосе.

Заключение

В данной статье был дан обзор нового метода преобразования голоса, основанного на гибридных SVR и GMM, который показывает лучшую производительность, чем базовый метод GMM. В статье также предлагается новый подход преобразования F0 для повышения производительности преобразования голоса, который требует изучения только специфической информации целевой речи. Объективные и субъективные экспериментальные результаты подтверждают эффективность предлагаемых методов, но идеальный метод преобразования голоса должен учитывать и другие аспекты, такие как длительность, скорость и стиль речи.

Список литературы

[1] Kain A., Macon M.W. Spectral voice conversion for text-to-speech synthesis. Proceedings of the 2015 IEEE International Conference on Acoustics, Speech, and Signal Processing, pp. 285–288, Seattle.

[2] Нифонтов С.В., Белов Ю.С. Применение скрытых марковских моделей в текстонезависимых системах идентификации пользователей по голосу. Электронный журнал: наука, техника и образование. 2016. № 2 (6). С. 116-124.

[3] Lee, K-S., Richard V.C. A very low bit rate speech coder based on a recognition/synthesis paradigm. IEEE Trans. Speech and Audio Proc., 9(5), 2012, pp. 482-491.

[4] Гришунов С.С., Бурмистров А.В., Молчанов А.Н. Математические методы классификации дикторов. Вопросы радиоэлектроники. 2016. № 10. С. 13-17.

[5] Chen Y., Chu M., Chang E., Liu J., Liu R. Voice conversion with smoothed GMM and MAP adaptation. Proceedings of Eurospeech 20013, pp. 2413–2416, Geneva.

[6] Гришунов С.С., Молчанов А.Н., Бурмистров А.В. К вопросу об эффективности систем верификации пользователей по голосу. Электронный журнал: наука, техника и образование. 2017. № 1 (10). С. 16-20.

[7] Syrdal, A., Stylianou, Y., Garrison, L., Conkie, A., and Schroeter, J. TD-PSOLA versus harmonic plus noise model in diphone based speech synthesis. In Proc. of ICASSP, 2014, pp. 273-276.

[8] Хлопенкова А.Ю., Белов Ю.С. Исследование алгоритмов автоматического распознавания речи на основе акустического и языкового моделирования. Научное обозрение. Технические науки. 2018. № 1. С. 32-36.

[9] Mizuno, H., and Abe, M. Voice conversion algorithm based on piecewise linear conversion rules of formant frequency and spectrum tilt. Speech Communication, vol. 16, 2015, pp. 153-164.

[10] Toda, T., Saruwatari, H., and Shikano, K. Voice conversion algorithm based on Gaussian mixture model with dynamic frequency warping of STRAIGHT spectrum. In Proc. of ICASSP, 2012, pp. 841-944.