УДК 621.391.037.372

ГИСТОГРАММНАЯ НОРМАЛИЗАЦИЯ РЕЧЕВЫХ ПРИЗНАКОВ В ЗАДАЧЕ ВЕРИФИКАЦИИ ДИКТОРОВ

Ю.Н. Матвеев, А.К. Шулипа

Содержится краткое описание алгоритма гистограммной нормализации речевых признаков применительно к задаче верификации дикторов. Приведены результаты верификационных тестов при различных параметрах и режимах нормализации. На основании полученных данных сделаны выводы об эффективности использования нормализации речевых признаков для улучшения качества верификации дикторов и найдены оптимальные условия использования алгоритма нормализации.

Ключевые слова: верификация дикторов, речевые признаки, гистограммная нормализация.

Введение

Как отмечено в работе [1], успех в распознавании личности по голосу (распознавании диктора) в гораздо большей степени зависит от метода компенсации канала, чем от выбора признаков. В качестве одного из путей решения проблемы канала в работе [2] было предложено использовать методы нормализации признаков путем их трансформирования (feature warping). Данный метод с 80-х годов прошлого века широко используется в обработке изображений [3–5], а для обработки речевых сигналов он получил распространение только в начале настоящего века [6].

Процедура нормализации речевых признаков выполняется на стадии предобработки речевого сигнала и заключается в приведении формы распределения признаков к заданному виду. В качестве заданного обычно используется нормальное (гауссово) распределение [6] и такую процедуру нормализации часто называют «гауссинизацией» [7]. Целью нормализации является уменьшение влияния вариативности внешних факторов на оценки параметров статистической модели голоса диктора и компенсация рассогласования каналов (условий тестирования и обучения). Таким образом, система верификации дикторов, использующая нормализованные признаки становится более робастной.

В случае дискретных сигналов нормализация основывается на модификации гистограмм распределений и поэтому называется гистограммной [8].

Известные методы гистограммной нормализации

На практике нормализация может быть выполнена одновременно на всем произнесении или последовательно, при смещении скользящего окна по речевому сигналу. В первом случае каждый элемент входных данных преобразуется с учетом его ранга, который определяется на всей последовательности векторов признаков. Во втором варианте нормализации производится трансформирование элемента, находящегося в центре окна, согласно его рангу среди других элементов из интервала, ограниченного окном. Алгоритм гистограммной нормализации речевых признаков с использованием скользящего окна позволяет выделять сигнал в зоне сильной нелинейности каналов связи, телефонных трубок.

В качестве речевых признаков для систем верификации наиболее часто рассматриваются вектора, которые состоят из набора мэл-частотных кепстральных коэффициентов (mel frequency cepstral coefficients, MFCC), а также их производных. Алгоритм нормализации признаков, предложенный в [2], предполагает нормализацию только MFCC-коэффициентов, на основе которых затем вычисляются оставшиеся компоненты вектора признаков (первые и вторые производные). Как показали результаты экспериментов, при таком исполнении алгоритм нормализации работает неэффективно.

Целью настоящей работы является разработка модификации алгоритма гистограммной нормализации, устраняющей недостатки известных решений.

Описание алгоритма гистограммной нормализации

В алгоритме гистограммной нормализации производится трансформирование вектора признаков в центре скользящего окна размерности N. Пусть известен входной набор векторов признаков в пределах окна $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\}$. Все векторы во входном наборе имеют $\dim(\mathbf{x})$ компонент. Далее необходимо вычислить относительный ранг для каждого i-го ($i \in \{1,\ldots, \dim(\mathbf{x})\}$) компонента вектора $x^{i\ /}_{\frac{N}{2}}$ в центре скользящего окна и найти для него соответствующее трансформированное значение $x^{i\ //}_{\frac{N}{2}}$. Скользящее окно последовательно сдвигается на вектор по всему произнесению и, таким образом, проводится трансформирование всех входных признаков, попавших в интервал $\left[\frac{N}{2}:1-\frac{N}{2}\right]$. Для произвольного положения скользящего окна ранг центрального вектора определяется согласно выражению

$$R\left(x^{i}_{\frac{N}{2}}\right) = \frac{1}{N}\sum_{k=1}^{N} l\left(x^{i}_{k}\right) - \frac{1}{2N}$$

при

$$l\left(x^{i}_{k}\right) = \begin{cases} 1, & x^{i}_{k} \le x^{i}_{\frac{N}{2}}; \\ 0, & x^{i}_{k} > x^{i}_{\frac{N}{2}}, \end{cases}$$

где $i \in \{1,\ldots, \dim(\mathbf{x})\}$ – индекс компоненты вектора признаков; $\dim(\mathbf{x})$ – число компонент вектора признаков $\mathbf{x}$; N – число векторов признаков в скользящем окне. Новое (трансформированное) значение признака из центра окна $x^{i\ //}_{\frac{N}{2}}$ находится по значению «старого» значения признака $x^{i\ /}_{\frac{N}{2}}$ из уравнения

$$R\left(x^{i\ /}_{\frac{N}{2}}\right) = Y\left(x^{i\ //}_{\frac{N}{2}}\right),$$

где

$$Y\left(x\right) = \int_{-\infty}^{x} \frac{e^{-\frac{t^2}{2}}}{\sqrt{2\pi}} dt\ .$$

Область значений величины R ограничена интервалом $R \in \left[\frac{1}{2N}:1-\frac{1}{2N}\right]$.

На практике, чтобы при каждом смещении окна избежать численного решения интегрального уравнения

$$Y_j = \int_{-\infty}^{x_j} \frac{e^{-\frac{t^2}{2}}}{\sqrt{2\pi}} dt\ ,$$

целесообразно предварительно рассчитать таблицу значений X_j для каждого Y_j следующим образом:

$$Y_j = \frac{1}{2N}(2j+1), \quad j \in [0\ldots N-1].$$

Точность вычисления интеграла для табличных значений Y_j должна быть не хуже $\frac{1}{2N}$.

Как было отмечено выше, нормализация выполняется только для векторов признаков, находящихся в центре скользящего окна. При этом часть векторов признаков, находящихся в интервалах длительностью $\frac{N}{2}$ в начале и конце произнесения, остается нетрансформированной. Для решения данной задачи необходимо рассчитать ранг этих векторов при начальном и конечном положении скользящего окна и выполнить для них трансформирование к нормальному распределению.

Рассмотренный алгоритм, аналогично известным алгоритмам нормализации кепстрального среднего и дисперсии [9], также дает набор центрированных кепстральных признаков MFCC с нулевым средним и единичной дисперсией.

Экспериментальные исследования предложенного метода

Эксперименты выполнялись для системы верификации дикторов на основе GMM-UBM-моделей голосов дикторов с использованием смесей гауссовых распределений (Gaussian mixture model, GMM) и универсальной фоновой модели (Universal Background Model, UBM) [1, 10]. Фонограммы для тестовых произнесений и эталонов выбирались из базы телефонных разговоров NIST-2008 [11]. Обучение проводилось на фонограммах телефонных баз NIST2005, NIST2006 [12, 13]. Все фонограммы, как для обучения, так и для тестирования, содержат речь только на английском языке. Условия обучения и тестирования при изменении параметров нормализации оставались постоянными. При обучении UBM использовались фонограммы 207 мужских и 275 женских голосов дикторов (3571 файлов). Множество обучения матрицы собственных каналов состояло из 190 мужских и 180 женских фонограмм (2583 файлов). Число компонент гауссовой смеси UBM $M = 512$, число собственных каналов $R = 50$. В качестве речевых признаков рассматривались MFCC-коэффициенты, их первая и вторая производные, объединенные в вектор размерности $L = 39$ (13mfcc + 13 delta mfcc + 13 delta delta mfcc). Выделение речевых сегментов проводилось с помощью детектора основного тона. Для построения GMM-эталона и компенсации влияния каналов использовалась схема Фогта [14]:

$$\boldsymbol{\mu} = \mathbf{m} + \mathbf{D}\mathbf{z} + \mathbf{U}\mathbf{x},$$

$$\boldsymbol{\mu}_0 = \mathbf{m} + \mathbf{D}\mathbf{z},$$

где $\mathbf{m}$ – супервектор (объединение векторов средних значений гауссоид смеси) GMM-UBM-размерности $P=M \cdot L$; $\mathbf{D}$ – диагональная матрица размерности $P \times P$; $\mathbf{U}$ – прямоугольная матрица собственных каналов размерности $R \times P$; $\boldsymbol{\mu}$ – супервектор GMM эталона размерности P; $\boldsymbol{\mu}_0$ – супервектор GMM эталона с учетом компенсации влияния эффектов канала размерности P; $\mathbf{z}$ – вектор факторов внутридикторской вариативности размерности P; $\mathbf{x}$ – вектор факторов вариативности, обусловленных эффектами канала размерности R.

Тестирование проводилось для фонограмм как с мужскими, так и с женскими голосами, причем пол дикторов на тестовых и эталонных фонограммах всегда совпадал. Число тестовых попыток для каждого типа сравнения представлено в табл. 1.

Тип сравнения	Женщины	Мужчины
Свой–свой	960	1274
Свой–чужой	13776	12672

Табл. 1. Число тестовых попыток при различных типах сравнения

Первоначально было исследовано влияние способов нормализации на качество верификации. Рассматривались два варианта нормализации: нормализация признаков путем вычитания кепстрального среднего (CMS) и нормализация трансформированием признаков (Feature Warping). При этом нормализация трансформированием признаков в одном случае выполнялась отдельно для каждой компоненты вектора признаков FW1, в другом – проводилась нормализация только для кепстральных коэффициентов вектора признаков и уже по ним рассчитывались производные первого и второго порядка FW2 [2]. Размер скользящего окна нормализации для FW1 и FW2 составлял 300 векторов. Результаты тестов верификации представлены в табл. 2. Для оценки качества верификации использовалось значение равновероятной ошибки (Equal Error Rate, EER) пропуска чужого и отклонения своего диктора.

	EER, %		
	CMS	FW1	FW2
Женщины	6,0	4,3	5,3
Мужчины	4,3	3,8	4,8

Табл. 2. Результаты верификации при использовании различных способов нормализации входных признаков

Далее было исследовано влияние размера скользящего окна нормализации на ошибку верификации дикторов. Результаты тестирования были получены при нормализации всех компонент вектора признаков как для MFCC-коэффициентов, так и для их производных (табл. 3).

Пол дикторов	Размер скользящего окна				
	100	200	300	400	500
Женщины	5,3	4,4	4,3	4,5	4,9
Мужчины	3,8	3,8	3,8	3,4	3,7

Табл. 3. Влияние размера скользящего окна нормализации на ошибку верификации дикторов (EER, %)

Заключение

В работе предложен алгоритм гистограммной нормализации кепстральных признаков со скользящим окном. Нормализация осуществляется независимо для всех компонент вектора кепстральных признаков – как для MFCC-коэффициентов, так и их первых и вторых производных.

Проведенные эксперименты показали, что использование предложенного алгоритма гистограммной нормализации существенно улучшает результаты верификации дикторов по сравнению с нормализацией методом вычитания кепстрального среднего. Это объясняется тем, что приведение формы распределения каждой компоненты вектора речевых признаков к заданному виду (нормальному распределению) позволяет устранить эффекты, связанные с рассогласованием условий обучения и тестирования.

Как показали эксперименты, нормализация отдельно каждой компоненты вектора кепстральных признаков является оптимальной и дает относительное снижение ошибки верификации более чем на 20% для фонограмм как с мужскими, так и с женскими голосами, при размере скользящего окна нормализации 400 векторов (4 с). Повышение робастности метода на более коротких произнесениях (1–4 с) является предметом дальнейших исследований.

Литература

1. Reynolds D., Rose R. Robust text-independent speaker identification using Gaussian mixture speaker models // IEEE Trans. Speech Audio Process. – 1995. – V. 3. – P. 72–83.
2. Pelecanos J., Sridharan S. Feature warping for robust speaker verification // Proc. A Speaker Odyssey. The Speaker Recognition Workshop, 2001. – P. 243–248.
3. Матвеев Ю.Н., Очин Е.Ф. Нелинейное преобразование видеосигнала на основе алгоритма скользящей эквализации гистограмм // Изв. вузов СССР. Радиоэлектроника. – 1985. – № 1. – С. 81–82.
4. Матвеев Ю.Н., Очин Е.Ф. Выполнение операции скользящего выравнивания гистограммы в матричном процессоре // Автометрия. – 1988. – № 1. – С. 14–17.
5. Кучеренко К.И., Матвеев Ю.Н., Очин Е.Ф. Устройство для скользящей эквализации гистограмм. Авт. свид. СССР, кл. G 06 F 15/36, 15/62. 1989.
6. Benesty J., Sondhi M.M., Huang Y. Springer Handbook of Speech Processing. – Springer, 2007. – P. 657–660.
7. Ramaswamy G.N., Gopinath R.A. Short-Time Gaussianization For Robust Speaker Verification // Acoustics, Speech, and Signal Processing. – 2002. – V. 1. – P. 681–684.
8. Матвеев Ю.Н., Очин Е.Ф. Структура устройства модификации гистограмм изображений // Тезисы докладов II Всесоюзной конференции «Методы и средства обработки сложной графической информации». – Горький: ГГУ, 1985. – 320 с.
9. Recent Advances in Robust Speech Recognition Technology / Javier Ramírez and Juan Manuel Górriz, eds. – Bentham Science Publishers, 2011. – 223 p.
10. Kenny P., Ouellet P., Dehak N., Gupta V., Dumouchel P. A Study of Inter-Speaker Variability in Speaker Verification // IEEE Trans. Audio Speech and Language Processing. – 2008. – V. 16. – № 5. – P. 980–988.
11. 2008 NIST Speaker Recognition Evaluation Test Set. [Электронный ресурс]. – Режим доступа: http://www.ldc.upenn.edu/Catalog/catalogEntry.jsp?catalogId=LDC2011S08, свободный. Яз. англ. (дата обращения 18.10.2012).
12. 2005 NIST Speaker Recognition Evaluation Training Data [Электронный ресурс]. – Режим доступа: http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2011S01, свободный. Яз. англ. (дата обращения 18.10.2012).
13. 2006 NIST Speaker Recognition Evaluation Training Set [Электронный ресурс]. – Режим доступа: http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2011S09, свободный. Яз. англ. (дата обращения 18.10.2012).
14. Vogt R., Sridharan S. Explicit Modelling of Session Variability for Speaker Verification // Computer Speech & Language, 2008. – V. 22. – № 1. – P. 17–38.

Матвеев Юрий Николаевич – Санкт-Петербургский национальный исследовательский университет информационных технологий, механики и оптики, ООО «ЦРТ-инновации», доктор технических наук, профессор, главный научный сотрудник, matveev@speechpro.com

Шулипа Андрей Константинович – ООО «ЦРТ-инновации», научный сотрудник, shulipa@speechpro.com