OSTIS-2015
(Open Semantic Technologies for Intelligent Systems)

УДК 004.822:514

ПРИМЕНЕНИЕ МЕТОДОВ ИНТЕЛЛЕКТУАЛЬНОГО АНАЛИЗА ВРЕМЕННЫХ РЯДОВ ДЛЯ ЗАДАЧИ КЛАСТЕРИЗАЦИИ ПОЛЬЗОВАТЕЛЕЙ ПО ГОЛОСУ

Вагин В.Н., Ганишев В.А.

Национальный исследовательский университет «МЭИ»,
г. Москва, Российская федерация

vagin@appmat.ru

v.ganishev@gmail.com

В данной работе рассматривается применение методов интеллектуального анализа временных рядов для задачи кластеризации пользователей по голосу. В качестве модели пользователя используется набор мел-частотных кепстральных коэффициентов. Использование метода нейросетевого сжатия данных позволяет сократить размерность вектора признаков. Кластеризация выполняется с помощью самоорганизующихся карт Кохонена.
Ключевые слова: кластеризация, временные ряды, мел-частотные кепстральные коэффициенты, самоорганизующиеся карты

Введение

Кластеризация пользователей по голосу – автоматическое соотношение каждой голосовой записи из некоторого набора к отдельному, как правило, анонимному пользователю. Чаще всего процесс кластеризации проходит без непосредственного контроля. Этот процесс часто является составной частью задач распознавания пользователей по голосу и распознаванию речи.

При решении данной задачи ставится вопрос не “что было сказано?”, а “кто это сказал?”. Кластеризация пользователей по голосу находит применение при анализе теле- и радио-трансляций, записей конференций и телефонных переговоров. Создание отдельной модели каждого пользователя является весьма ресурсоемким, когда речь идет о массовых событиях и системах, содержащих тысячи записей сотен пользователей: системы в таком случае смогут работать лишь на предопределенном множестве пользователей, что лишает их гибкости. С другой стороны, создание отдельной модели для пользователя, редко использующего эти системы, является неоправданным с точки зрения экономии ресурсов.

В последние годы решение данной задачи является одним из приоритетных направлений таких областей исследований, как анализ сигналов, компьютерная безопасность и искусственный интеллект.

Для решения этой задачи часто применяются следующие модели:

- смеси гауссовских распределений (*GMM*, *Gaussian Mixture Model*) [Han et al., 2008];
- скрытые марковские модели (*HMM*, *Hidden Markov Model*) [Ajmera et al., 2003];
- гистограммные модели (*Histogram Model*) [Rodriguez Fuentes et al., 2004];
- алгоритм спектральной кластеризации Ына-Джордана-Вайса (*Ng-Jordan-Weiss spectral clustering algorithm*)[Ning et al., 2004];
- алгоритм байесовской адаптации [Faltlhauser et al., 2006]

Считается, что данные модели более подходят для описания поведения, характерного именно для голосового сигнала. Тем не менее, перспективным представляется применение инструментов интеллектуального анализа временных рядов.

В данной работе предлагается использовать другой подход, основанный на применении классического алгоритма кластеризации для характеристик, выделенных из сигнала и уникальных для каждого пользователя. В качестве таких характеристик используются мел-частотные кепстральные коэффициенты.

1. Модель пользователя

1.1. Причины использование мел-частотных кепстральных коэффициентов

Звуковой сигнал является одним из средств взаимодействия человека с окружающей средой и людей между собой. Голос зависит от многих физиологических параметров говорящего и является по своей сути индивидуальной характеристикой каждого человека. Тем не менее, голос не является постоянной характеристикой, он изменяется в течение жизни человека, на него также влияют состояние здоровья и эмоции.

Современные средства записи позволяют представить звуковой сигнал в виде временного ряда, показывающего изменение частоты во времени. Спектр сигнала, его представление в частотном пространстве является более информативным для анализа, чем сигнал сам по себе. Для вычисления спектра часто используется быстрое преобразование Фурье, алгоритм которого является достаточно простым для реализации и имеет сложность $O(N\log_2 N)$, меньшую, чем сложность классического алгоритма дискретного преобразования Фурье $O(N^2)$ [Cooley et al, 1965]. Люди реагируют на частотные изменения, поэтому при решении задач, связанных с анализом человеческого голоса, часто используют «кепстр» (*cepstrum*) [Bogert et al., 1963] – результат применения преобразования Фурье к спектру сигнала.

Также в процессе эволюции звуки в более низком частотном диапазоне содержали в себе больше полезной информации, чем находящиеся в более высоком частотном диапазоне. С учетом этих особенностей человеческого слуха были разработаны мел-частотные кепстральные коэффициенты («мел» является сокращением английского слова «melody» - мелодия) [Vyas et al., 2013]. С помощью данных коэффициентов более тщательно анализируется информация, получаемая из низкочастотного диапазона, а влияние высокочастотных составляющих, обычно содержащих посторонний шум, на результат распознавания уменьшается.

Вся голосовая запись разделяется на небольшие интервалы, длительностью ~10-30мс (время квазистационарности сигнала), называемые фреймами. Для каждого фрейма отдельно рассчитывается набор мел-частотных кепстральных коэффициентов, который в дальнейшем будет использоваться для кластеризации.

1.2. Вычисление мел-частотных кепстальных коэффициентов

Алгоритм вычисления мел-частотных кепстральных коэффициентов можно разбить на следующие этапы [Molau et al., 2001]:

а. разбивание сигнала на фреймы;

б. применение весовой функции (окна) к каждому фрейму;

в. применение преобразования Фурье;

г. использование мел-частотного фильтра;

д. вычисление кепстра.

а. Разбиение сигнала на фреймы

Звуковой сигнал в общем случае не является стационарным, т.е. их амплитуда и спектр изменяются во времени, что приводит к невозможности применения многих техник анализа. Но отдельно взятый короткий интервал порядка 10-30мс можно считать стационарным. Часто применяют следующую методику деления сигнала на фреймы: сигнал разделяется на интервалы длиною N мс следующим образом: начало первого фрейма совпадает с началом записи, второй фрейм начинается через M мс интервалов (M < N), соответственно он на N-M мс перекрывает первый фрейм. На Рис. 1 показан случай для N = 20 мс и M = 16 мс.

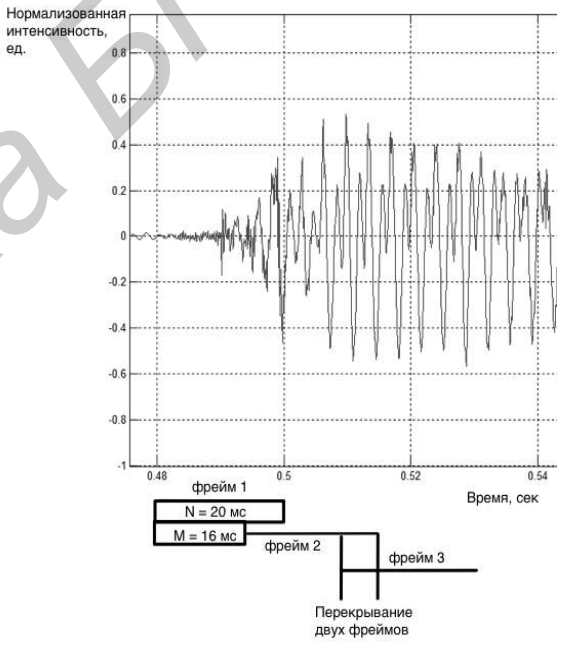

Рисунок 1 – Пример разбиения на фреймы

Несмотря на стационарность, такое представление сигнала не позволяет использовать преобразование Фурье. Если частоты гармоник (частотных составляющих) сигнала не совпадают с базисными частотами преобразования Фурье, то в спектре будут возникать «лишние» гармоники, которые будут лишь «зашумлять» полученное представление. Данный эффект носит название «размытие спектра» или «спектральная утечка».

б. Применение весовой функции (окна)

Одним из возможных вариантов решения возникшей проблемы является применение к сигналу весовой функции специального вида:

$$\omega(n), 0 \le n \le N-1 \qquad (1)$$

Результат применения весовой функции к каждому фрейму выглядит следующим образом (Рис. 2а):

$$y(n) = x(n) \cdot \omega(n), 0 \le n \le N-1 \qquad (2)$$

где *x(n)* – значение временного ряда в точке *n*, а *y(n)* – взвешенное значение временного ряда в точке *n*.

Наиболее предпочтительным является применение «мягких» весовых функций, которые сводят значения на границах фрейма к нулю. Эта операция называется «сглаживанием». Наиболее часто используемой является весовая функция Хэмминга, которую можно представить следующей формулой [Bhatnagar et al., 2012]:

$$\omega(n) = 0.53836 - 0.46165 \cdot cos\left(\frac{2\pi n}{N-1}\right) \qquad (3)$$

Преобразование Фурье примененное к «взвешенному» временному ряду дает более четкий спектр (Рис. 2б).

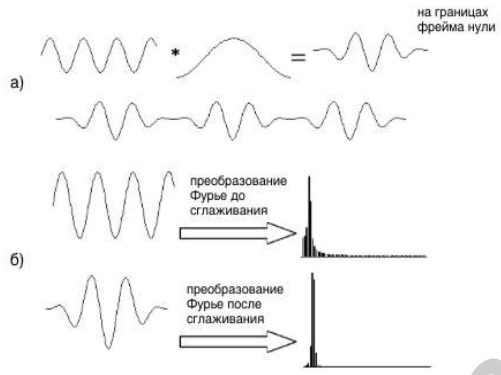

Рисунок 2 – Сглаживание сигнала:
а) Применение весовой функции к фрейму
б) Применение преобразование Фурье

в. Преобразование Фурье

На следующем этапе необходимо применить преобразование Фурье, которое переведет сигнал из временного пространства в частотное. На практике чаще всего применяется быстрое преобразование Фурье, который имеет следующий вид [Cooley et al, 1965]:

$$Y_n = \sum_{k=0}^{N-1} y_k \cdot e^{\frac{-2\pi jkn}{N}}, 0 \le n \le N-1, j = \sqrt{-1} \qquad (4)$$

где y_k – взвешенное значение временного ряда в точке k,

Y_n – комплексная амплитуда n-той гармоники сигнала, представляемого временным рядом.

Результатом данного этапа является спектр сигнала.

г. Использование мел-частотного фильтра

На данном этапе к спектру сигнала применяется специального вида фильтр. Каждому значению частоты, полученному на предыдущем шаге ставится в соответствие значение на мел-частотной шкале. Значения данной шкалы для частот ниже 1000 Гц точно соответствуют спектру сигнала, полученному при преобразовании Фурье, частоты выше 1000 Гц – логарифмируются. В результате получается модифицированный энергетический спектр сигнала *mel(f)* для каждой гармоники частоты *f*, для вычисления которого используется следующая приближенная формула [Molau et al., 2001]:

$$mel(f) = 2595 \cdot lg\left(1 + \frac{f}{700}\right) \qquad (5)$$

К данному спектру применяется фильтр специального вида, ставящий в соответствие каждой частоте определенный набор мел-коэффициентов $\tilde{S}_k$, 1,...,K, где K – количество мел-коэффициентов, на практике часто выбирают значение от 12 до 24.

д. Использование мел-частотного фильтра

На предидущем шаге алгоритма полученные коэффициенты $\tilde{S}_k$ необходимо перевести в мел-кепстальное пространство. Для этого удобно использовать дискретное косинусоидальное преобразование, которое описывается следующей формулой [Chen et al., 1977]:

$$\tilde{C}_n = \sum_{k=1}^{K} lg(\tilde{S}_k) \cdot cos\left[n\left(k - \frac{1}{2}\right)\frac{\pi}{K}\right], 0 \le n \le K \qquad (6)$$

где $\tilde{C}_n$ – полученные мел-частотные кепстральные коэффициенты.

1.3. Акустические векторы

Данный алгоритм применяется к каждому фрейму, в результате чего последнему соответствует набор мел-коэффициентов, который используется в большинстве работ как модель пользователя для кластеризации и называется акустическим вектором.

Но изменение мел-коэффициентов также содержит определенную информацию о пользователе. Основным отличием данной работы от предыдущих является расширение акустического вектора путем учета динамики изменения мел-коэффициентов δ_i, которая выражается разностью мел-частотных кепстральных коэффициентов, данного фрейма и предыдущего:

$$\delta_i(\tilde{C}_k[i]) = \tilde{C}_k[i-1] - \tilde{C}_k[i] \qquad (7)$$

Данный подход позволяет анализировать изменения мел-частотных кепстральных коэффициентов, что также несет в себе информацию, идентифицирующую пользователя.

При данном подходе первый фрейм не может использоваться для кластеризации, так как изменение мел-частотных кепстральных коэффициентов будет нулевым. А L – количество элементов акустического вектора x – увеличивается вдвое: $L = |x| = \left|\left[\tilde{C}_1, \dots, \tilde{C}_k, \delta(\tilde{C}_1), \dots, \delta(\tilde{C}_k)\right]\right| = 2 \cdot K$.

2. Применение метода нейросетевого сжатия для акустических векторов

Большая размерность акустического вектора является проблемой, которая снижает скорость кластеризации. Для сокращения размерности акустического вектора предлагается использовать метод нейросетевого сжатия данных.

Данный метод реализуется с помощью трехслойной нейронной сети следующего вида [Tishby et al., 1999]:

- входной и выходной слои сети идентичны и соответствуют набору элементов акустического вектора;
- скрытый слой содержит меньшее количество нейронов.

После обучения сети коэффициенты скрытого слоя будут представлять собой новый акустический вектор меньшей размерности. В общем случае данный алгоритм может применяться несколько раз для дальнейшего сокращения размерности до тех пор, пока среднеквадратичная ошибка E не превышает заранее заданный порог:

$$E = \sum_{k=1}^{n} E(k) = \frac{1}{2}\sum_{k=1}^{L}(y_k - x_k)^2 < \varepsilon \qquad (8)$$

где E(k) – среднеквадратичная ошибка для k-го акустического вектора записи;

L – количество элементов акустического вектора;

y_k – полученное значение акустического вектора на выходе нейронной сети;

x_k – значение акустического вектора на входе нейронной сети;

ε – некоторый заранее определенный порог.

3. Самоорганизующиеся карты

Задача определения пользователя по акустическому вектору относится к классу задач распознавания по шаблону. Использование нейронных сетей Кохонена для решения задачи кластеризации пользователей по голосу было выбрано в силу точности кластеризации и ее скорости [Mori et al., 2001].

В данной работе используются самоорганизующаяся карта Кохонена [Kohonen, 1995]. Она представляет собой нейронную сеть с двумя слоями, причем нейроны первого (распределительного) слоя соединены со всеми нейронами второго (выходного) слоя, которые расположены в виде двумерной решетки. Количество нейроном в выходном слое определяет максимальное количество групп, на которые система может разделить входные данные.

Для обучения сети Кохонена используется соревновательный метод [Kohonen, 1995]. На каждом шаге обучения из исходного набора данных случайно выбирается один вектор. Затем производится поиск нейрона выходного слоя, для которого расстояние между его вектором весов и входным вектором - минимально.

Алгоритм обучения сети Кохонена выглядит следующим образом [Головко, 2001]:

1. Инициализация малыми случайными значениями на отрезке [-1, 1] матрицы весов сети W размерности $L \times T$, где T – количество записей, которые необходимо кластеризовать;
2. организация акустических векторов в очередь в случайном порядке. Все вектора помечены как необработанные;
3. выбор первого необработанного элемента x из очереди;
4. для каждого выхода сети j вычисляются расстояния d_j между его вектором весов w_j и входным акустическим вектором. В данной работе используется квадрат евклидова расстояния:

$$d_j = \rho(w_j, x), \qquad (9)$$

где

$$\rho(w_j, x) = \sum_{i=1}^{n}(w_{ji} - x_i)^2 \qquad (10)$$

5. поиск выходного нейрона j_m с минимальным расстоянием d_{jm}:

$$j_m = \arg\ min\ (d_j) \qquad (11)$$

6. вычисление изменения весов $\Delta W = \{\Delta w_j\}$ для всех нейронов j выходного слоя:

$$\Delta w_j = (w_j - x) \cdot h(u, c, t) \cdot \eta \qquad (12)$$

где η – коэффициент скорости обучения;

c – номер нейрона победителя j_m в двумерной решетке второго слоя;

j – номер нейрона в двумерной решетке второго слоя;

w_j – вектор весовых коэффициентов связи входного слоя и нейрона с номером j;

x – акустический вектор на входе сети;

$h(u,c,t)$ – функция окрестности. В данной работе использована функция Гаусса:

$$h(u, c, t) = e^{\frac{\rho(c,u)}{\sigma(t)}} \qquad (13)$$

где t – параметр времени;

σ – радиус окрестности h:

$$\sigma(t) = \frac{1}{e^{t^{-2}}} \qquad (14)$$

7. корректировка матрицы весов W нейронной сети:

$$W := W - \Delta W \qquad (15)$$

8. элемент x входной очереди помечается как обработанный;
9. если в очереди имеются акустические векторы, помеченные как необработанные, то переход к п.3;
10. если критерий останова обучения не достигнут, то переход к п.2. В данной работе в

качестве критерия останова используется проверка стабилизации выходов сети: когда акустические векторы на последующих этапах обучения перестают переходить между кластерными элементами. Математически это выражается в том, что определитель матрицы изменений весов меньше некоторого порогового значения.

11. окончание алгоритма.

После применения данного алгоритма одному пользователю будет соответствовать несколько кластеров. Считаем, что голосовой сигнал принадлежит пользователю в том случае, если более половины фреймов были ассоциированы с этим пользователем. На Рис.3 показан результат кластеризации для двух пользователей. Акустические вектора первого пользователя обозначены кругом, второго – треугольником. Сверху представлены акустические вектора одной записи. На изображении видно, что они ассоциированы с разными кластерами.

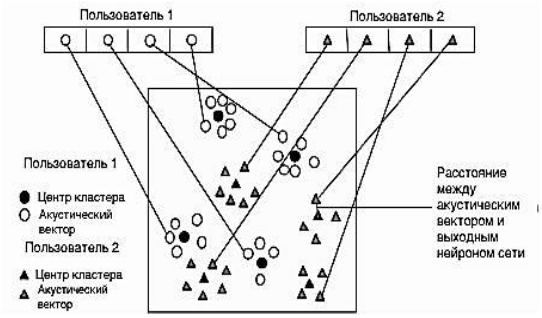

Рисунок 3 – Пример результата кластеризации

4. Практическая реализация данного метода

Для вычисления мел-частотных кепстральных коэффициентов используются средства свободно распространяемого фреймворка Sphinx 4, разработанной в университете Карнеги-Меллон [Walker et al.,2004]. Данный комплекс реализует множество функций, необходимых для распознавания пользователей по голосу, и обладает простым интерфейсом.

Как уже было сказано ранее, большая размерность акустического вектора является проблемой, поэтому для ее решения был предложен метод нейросетевого сжатия.

Программный комплекс имеет следующую структуру:

- Блок преобразования входного голосового сигнала. На данном этапе входной голосовой сигнал в формате .wav приводится к удобной для обработки форме представления в виде временного ряда.
- Блок обработки сигнала. На данном этапе на основе сигнала вычисляются мел-частотные кепстральные коэффициенты.
- Блок предобработки акустического вектора. В данном блоке реализуется сжатие акустического вектора до меньшего размера.
- Блок кластеризации. Здесь непосредственно применяется алгоритм кластеризации, описанный выше.
- Блок принятия решения. На данном этапе принимается решение о принадлежности записи определенному шаблону пользователя.

Система реализована в виде библиотеки C++, так как данный язык обладает кроссплатформенностью и высокой производительностью вычислений.

В настоящее время ведется исследование возможности усовершенствования алгоритма кластеризации для автоматического переобучения при добавлении голосовых записей новых пользователей.

Заключение

Данная работа рассматривает кластеризацию пользователей по голосу. В качестве характеристик пользователя предлагается использовать расширенный акустический вектор каждого фрейма голосовой записи, состоящий из мел-частотных кепстральных коэффициентов, а их изменения относительно прошлого фрейма.

Для сокращения размерности акустического вектора в данной работе предложен метод нейросетевого сжатия данных, который позволяет сократить размерность исходных данных для задачи кластеризации.

В качестве алгоритма кластеризации используются самоорганизующиеся карты Кохонена, так как использование нейронных сетей для кластеризации позволяет учитывать неочевидные закономерности в голосовых характеристиках, такие как характер изменений частоты голоса и т.д.

Предложенный метод учитывает дополнительные особенности голосовых характеристик каждого пользователя, такие как скорость изменения частот голоса.

Библиографический список

[Ajmera et al., 2003] Ajmera J. [et al.] A Robust Speaker Clustering Algorithm // IEEE Workshop on Automatic Speech Recognition and Understanding, 2003, 2003 , C. 411-416

[Bhatnagar et al., 2012] Bhatnagar A.C. [et al.] Analysis of Hamming window using advance peak windowing method // Interational Journal of Scientific Research Engineering&Technology (IJSRET) Vol.1 Issue 4, 2012, C. 15-20

[Bogert et al., 1963] Bogert B.P. [et al.] The Quefrency Alanysis of Time Series for Echoes: Cepstrum, Pseudo Autocovariance, Cross-Cepstrum and Saphe Cracking // Proceedings of the Symposium on Time Series Analysis (M. Rosenblatt, Ed) Chapter 15, New York: Wiley, 1963, C. 209-243

[Chen et al., 1977] Chen W.-H. [et al.] A Fast Computationl Algorithm for the Discrete Cosine Transform // IEEE Transactionsof Communications, Vol.Com-25, No.9, 1977, C.1004-1009

[Cooley et al, 1965] Cooley J.W. [et al.]An Algorithm for the Machine Calculation of Complex Fourier Series // Mathematics of Computation, 1965, C. 297-301

[Faltlhauser et al., 2001] Faltlhauser R. [et al.] Robust Speaker Clustering in Eigenspace // IEEE Workshop on Automatic Speech Recognition and Understanding, 2001, C. 57-60

[Han et al., 2008] Han K.J. [et al.] Agglomerative Hierarchical Speaker Clustering using Incremental Gaussian Mixture Cluster Modeling // Proceedings of InterSpeech, 2008, C. 20-23

[Kohonen, 1995] Kohonen T. Self-Organizing Maps // Springer, 1995

[Linde et al., 1980] Linde, Y. [et al.] An Algorithm for Vector Quantizer Design // IEEE Transactions on Communications 28, 1980, C. 84-95

[Molau et al., 2001] Molau S. [et al.] Computing mel-frequency cepstral coefficients on the power spectrum // IEEE International Conference on Accoustics, Speech, and Signal ProcessingVol.1, 2001, C.73-76

[Mori et al., 2001] Mori K. [et al.] Speaker change detection and speaker clustering using VQ distortion for broadcast news speech recognition // Acoustics, Speech, and Signal Processing, 2001. Proceedings. (ICASSP '01). 2001 IEEE International Conference on (Volume:1), 2001, C. 413-416

[Ning et al., 2006] Ning H. [et al.] A Spectral Clustering Approach to Speaker Diarization // Proc. ICSLP, 2006, C.

[Tishby et al., 1999] Tishby N., Pereira F., and Bialek W. The Information Bottleneck Method // The 37th annual Allerton Conference on Communication, Control, and Computing, 1999, c.368-379

[Rodriguez Fuentes, 2004] Rodriguez Fuentes L.J. [et al.] A Speaker Clustering Algorithm for Fast Speaker Adaptation in Continuous Speech Recognition // Text, Speech and Dialogue: Lecture Notes in Computer Science Volume 3206, 2004, C. 433-440

[Vyas et al., 2013] Vyas G. [et al.] Speaker Recognition System Based on MFCC and DCT // Internatonal Journal of Engineering and Advanced Technology(IJEAT) Vol. 2, Issue 5, 2013

[Walker et al.,, 1975] Walker W. [et al.] Sphinx-4: A flexible open source framework for speech recognition // Technical Report , 2004

[Вагин и др., 2008] Вагин В.Н. [и др.] Достоверный и правдоподобный вывод в интеллектуальных системах, 2-ое издание, исправленное и дополненное // – М. : ФИЗМАТЛИТ, пол ред. Вагина В.Н. и Поспелова Д.А., 2008, 712 с..

[Головко, 2001] Головко В.А. Нейронные сети: обучение, организация, применение // М. : ИПРЖР, 2001

APPLICATION OF TIME SERIES ANALYSIS FOR SPEAKER CLUSTERING

Vagin V.N., Ganishev V.A.
National Research University «MPEI», Moscow, Russia
vagin@appmat.ru
v.ganishev@gmail.com

The purpose of this paper is the introduction of time series analysis methods to the problem of speaker clustering. User's model used for clustering is based on the mel-frequency cepstral coefficients. We consider the use of methods of neuro-network data compression to reduce the dimensionality of the feature vector. Clustering is performed using self-organizing Kohonen maps.

Introduction

Speaker clustering is an automatic classification of voice recordings on some patterns of users, often without direct supervision. This process is often an integral part of the user recognition and speech recognition tasks.

By solving this problem the main question is not "What was said?" but "Who said that?". Speaker clustering is used in the analysis of television and radio broadcasts, recordings of telephone conversations and conferences. Creation of separate model for each user is very resource-intensive in context of mass events and systems and makes the system over-fitted and not flexible.

Main Part

In this paper it is proposed to use speaker model based on mel-frequency cepstral coefficients (MFCC). It views the input voice signal in terms of model, that is close to perception oh the human ear. It provides more detailed analysis of low-frequency component of the recording that contains more information needed for clustering.

The main contrast of this article from a huge variety of previous works is the usage not only the computed MFCC as features for clustering, but also the dynamic difference between MFCC of the considered frame and the previous.

The expansion of the acoustic vector follows to complicating and deceleration at the clustering phase. For solving this problem, the use of bottleneck method for feature compression is proposed.

The concept of self-organizing Kohonen maps is used for clustering. This method provides the clear and fast algorithm that is suitable for speaker clustering, because each user will have several clusters. That helps to overview different phonemes said by user in different situations.

Conclusion

This paper considers the problem of speaker clustering. As the features for clustering it is proposed to use advanced acoustic vector for each frame of a voice recording, consisting of mel-frequency cepstral coefficients and their dynamic changes over the previous frame.

To reduce the dimension of the acoustic vectors the bottleneck method for data compression in terms of a neural net is used.

The self-organizing Kohonen map is applied as a clustering algorithm.