УДК 612.382

РЕАЛИЗАЦИЯ И ВЫБОР ПАРАМЕТРОВ ПРИ ИСПОЛЬЗОВАНИИ АЛГОРИТМА ВЫРАВНИВАНИЯ ВРЕМЕННЫХ МАСШТАБОВ ДЛЯ СИСТЕМ КОНВЕРСИИ ГОЛОСА

ТХАЙ ЧУНГ КИЕН

Белорусский государственный университет информатики и радиоэлектроники
П. Бровки, 6, Минск, 220013, Беларусь

Поступила в редакцию 2 июня 2008

Алгоритм выравнивания временных масштабов (Dynamic Time Warping — DTW) широко применяется в распознавании речи. Целью DTW является нахождение функции выравнивания, которая минимизирует общее расстояние между соответствующими фреймами речевых сигналов. В статье рассматриваются алгоритм DTW и его применение в области конверсии голоса, где алгоритм DTW используется для совмещения речевых фреймов двух различных дикторов. Данный алгоритм служит предварительной обработкой и используется на этапе обучения, оказывая прямое влияние на точность функции конверсии. В данной работе предлагается способ выбора параметров тракта для нахождения совмещения. Алгоритм реализован как с использованием коэффициентов линейной спектральной частоты (Line spectral frequencies — LSF), так и с использованием кепстральных коэффициентов. Показано, что кепстральные коэффициенты дают наилучший результат, который исключает получение вырожденных матриц и упрощает этап обучения, а так же существенно повышает качество конверсии речи.

Ключевые слова: алгоритм выравнивания временных масштабов (DTW), конверсия голоса (VC), линейные спектральные частоты, кепстральный коэффициент.

Введение

Задача распознавания речи связана с определением длительности того или иного слова, поскольку любой диктор произносит одно и тоже слово с разной продолжительностью. Эта особенность является серьезным препятствием при распознавании. Система распознавания различает структуры двух различных типов: словесный модуль и подсловесный модуль [1]. При использовании структуры словесного модуля входная речь сравнивается с каждой словесной моделью (эталоном) для выбора словесного модуля с наименьшим расстоянием. В этом случае сравниваются сегменты сигнала большой продолжительности (отдельные слова). В отличие от словесного сравнения, подсловестные модули сравниваются при помощи вычисления расстояния между короткими фрагментами речи (десятки миллисекунд). Для сравнения модулей используется алгоритм DTW, который является самым популярным методом, компенсирующим разницу в скорости речи двух дикторов. Этот алгоритм также используется в задачах верификации диктора с применением в системе эталонной модели [2].

Задача конверсии голоса на этапе обучения требует однозначного соответствия между исходными и целевыми фреймами. Хотя большинство методов использует обучающие записи с одинаковым содержанием, произнесенные исходным и целевым дикторами, тем не менее, в них

необходимо найти соответствие на уровне фреймов. Реализации данного подхода предлагались в различных системах конверсии голоса.

Альтернативой DTW является ручное совмещение, что было использовано в некоторых параметрических системах. Например, в [3] форманты регулировались вручную для повышения точности системы. Несмотря на хорошую точность этого метода, для его реализации требуются колоссальные временные затраты, что часто делает невозможным его применение.

Часто в системах конверсии голоса используется скрытая марковская модель (HMM) [4] для регулирования акустических параметров и вектора LSF. Записываются опорные предложения, произнесенные исходным и целевым дикторами. Для каждого предложения извлекаются кепстральные коэффициенты из логарифма мощности и пересечения нулевого уровня для каждого анализируемого фрейма. На основе последовательности параметрических векторов выполняется обучение для каждого опорного предложения с использованием данных исходного диктора. Количество состояний для каждого предложения HMM пропорционально периоду произнесения. Затем последовательность наилучших состояний оценивается при помощи алгоритма Витерби. Вектор средних коэффициентов LSF для каждого состояния вычисляется для обоих, исходного и целевого, дикторов при использовании векторов фрейма, соответствующих номеру данного состояния. Наконец, данный вектор является совмещенным вектором исходного и целевого дикторов. Использование HMM показывает хорошие результаты, однако сопряжено с большими сложностями и не дает возможности получить точное взаимооднозначное отображение [4].

Алгоритм DTW вычисляет нелинейное отображение одного сигнала в другой с помощью минимизации расстояний между сигналами. Целью DTW является образование функции выравнивания, которая минимизирует расстояние между соответствующими фреймами. Два сигнала совмещаются и минимальное расстояние достигается с учетом накопленного расстояния. Алгоритм DTW может быть использован для определения поведения параметров во времени, которые описывают динамическую конфигурацию вокального тракта. В частности, DTW используется для нахождения временного соответствия так, чтобы минимизировать спектральное расстояние между фреймами исходного и целевого дикторов в системах конверсии голоса.

Алгоритм выравнивания временных масштабов (Dynamic time warping algorithm — DTW)

Предложим, что Q и C — две последовательности времени длиной n и m соответственно [5], где:

$$Q = q_1, q_2, ..., q_i, ..., q_n \tag{1}$$

$$C = c_1, c_2, ..., c_i, ..., c_m \tag{2}$$

Чтобы совмещать две эти последовательности при помощи DTW, составляется матрица размерностью $n \times m$, где ее i-й, j-й элементы есть расстояние $d(q_i, c_j)$ между 2-мя точками q_i и c_j. Расстояние $d(q_i, c_j)$ является евклидовым, которое определяется как

$$d\left(q_i, c_j\right) = \left(q_i - c_j\right)^2 \tag{3}$$

Отдельные элементы $d(q_i, c_j)$ называются локальными расстояниями. Каждый элемент (i, j) матрицы соответствует совмещению между точками q_i и c_j. Процесс выравнивания может быть очень эффективно реализован при помощи динамического программирования для оценки последующей рекуррентности, которая определяет накопленное расстояние $g(i, j)$. Расстояние $d(i, j)$ находится в настоящей секции и минимум накопленного расстояния между соседними элементами определяется как

$$g\left(i, j\right) = d\left(q_i, c_j\right) + \min\left\{g\left(i-1, j-1\right), g\left(i-1, j\right), g\left(i, j-1\right)\right\}. \tag{4}$$

Если w — траектория выравнивания, то она ограничена следующими условиями:

– граничное условие: $w_1=(1, 1)$ и $w_K=(m, n)$; требуемая траектория выравнивания такова, чтобы она начиналась первым и заканчивалась последним элементом главной диагонали;

– непрерывность: заданные $w_k=(a, b)$ и $w_{k-1}=(a', b')$, где $a–a'\leq 1$ и $b–b'\leq 1$. Что является ограничением разрешенной траектории выравнивания для соседних элементов.

– монотонность: заданые $w_k=(a, b)$ и $w_{k-1}=(a', b')$, где $a–a'\geq 0$ и $b–b'\geq 0$. Это требование монотонного увеличения во времени точек W. Например, нельзя отображать точку 6, точку 4, а затем точку 5 или точку 6, точку 5, а затем опять точку 6.

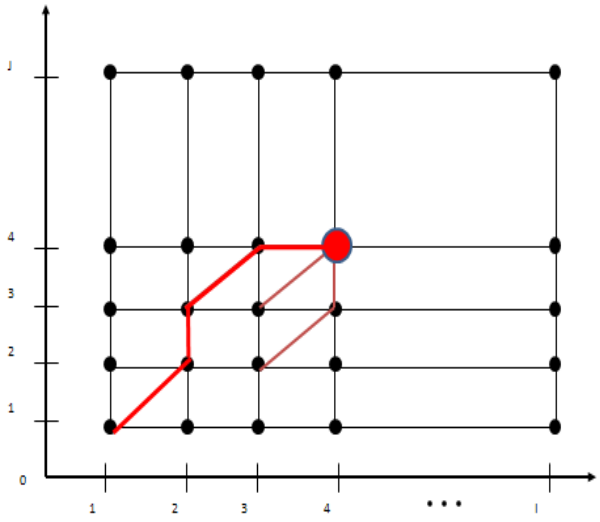

Рис. 1. Иллюстрация алгоритма выравнивания временных масштабов

Выбор параметров и экспериментальные результаты

Как было отмечено в [6], коэффициенты LPC используются для представления параметров вокального тракта. Параметры LPC образуют перцепционное описание спектральной огибающей, так как они описывают перцепционно важные спектральные пики более точно, чем точки минимума спектра. Эти параметры также широко используются в кодировании речи. Однако они имеют следующие недостатки:

– множество коэффициентов LPC не достаточно однородно;

– коэффициенты LPC плохо квантуются, т.е. малая ошибка квантования может приводить к значительному искажению спектра,

– коэффициенты LPC также плохо интерполируются, т.е. нельзя, имея два LPC вектора, полученные в разные моменты времени, с удовлетворительной точностью предсказать значения вектора, находящегося между ними.

Чтобы исключить эти недостатки, были предложены модификации LPC. Например: коэффициенты отражения, кепстральные коэффициенты, логарифм отношения площади (Log Area Ratios — LAR). Наряду с этим, для кодирования и конверсии речи, широко используется другое представление LPC — спектральные линии (Line Spectrum Frequencies — LSF), которое было предложено в [1, гл. 5 с. 83–129]. LSF имеет следующие полезные свойства:

простая проверка устойчивости. Если LSF расположены по возрастанию в интервале [0:5], то соответствующий фильтр гарантированно устойчивый;

можно делать интерполяцию;

LSF сильно коррелированны между собой, они могут быть эффективно квантованы;

когда два значения LSF близки друг к другу, спектральный пик находится между ними, это полезно для слежения за формантами и спектральными пиками.

Расстояние между двумя фреймами исходного и целевого дикторов определяется как [7]

$$d\left(A,B\right)=\sqrt{\frac{1}{p}\sum_{i=1}^{p}\left(L_A^{m,i}-L_B^{m,i}\right)^2}\,, \tag{5}$$

где p — порядок LPC, $L^{m,i}$ — i-я LSF m-го фрейма.

Кепстральное расстояние представляет другую меру параметров LPC. Коэффициенты LPC также могут использоваться для вычисления кепстральных коэффициентов полной разницы между кепстрами исходного и соответствующего закодированного голоса [8]. В отличие от непосредственного вычисления из формы волны голоса, кепстр получается из коэффициентов LPC, который является гладким речевым спектром [9, гл. 3, с. 91–94]:

$$\begin{aligned} c\left(n\right)&=\sum_{k=1}^{n-1}\left(1-\frac{k}{n}a_k\right)c\left(n-k\right)+a_n, && 1<n\le p;\\ c\left(n\right)&=\sum_{k=1}^{n-1}\left(1-\frac{k}{n}a_k\right)c\left(n-k\right), && n>p;\\ c\left(1\right)&=a_1. \end{aligned} \tag{6}$$

Кепстральное расстояние вычисляется по формуле (5), где $L^{m,\mathrm{i}}$ — i-й кепстральный параметр m-го фрейма.

Чтобы проиллюстрировать выполнение алгоритма и отличие использования двух типов параметров (LSF и кепстр), два предложения были совмещены и показаны. Рис. 2 показывает траектории мужской и женской речи. Как видно из рисунка, продолжительности (число фрейм) двух предложений различны.

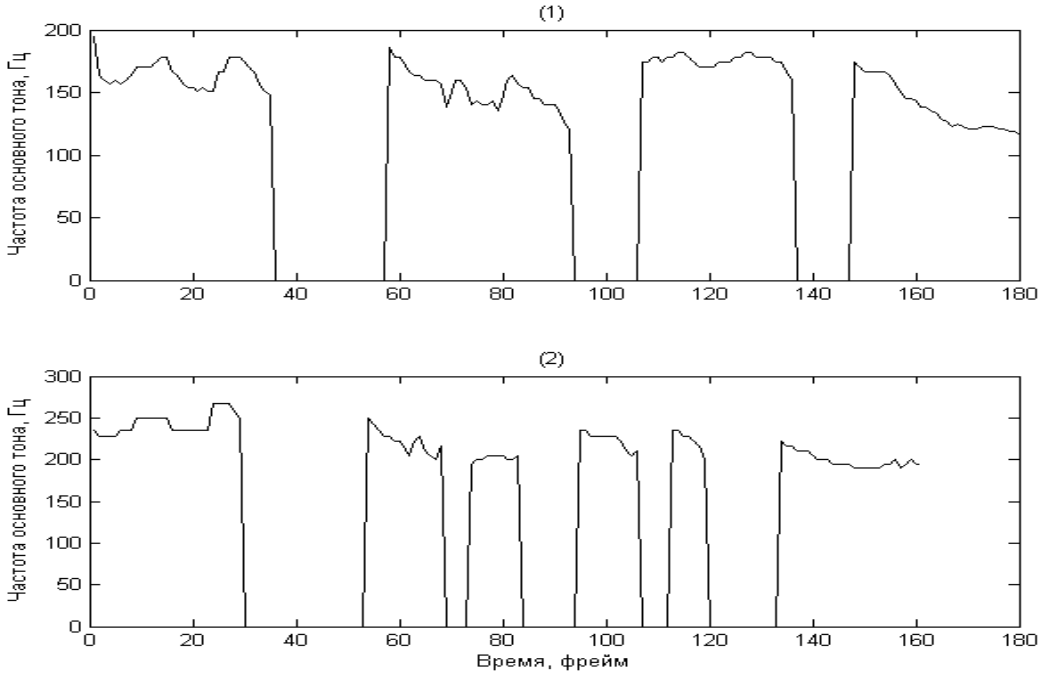

Рис. 2. Траектория частоты основного тона: *1* — исходная речь (мужская); *2* — целевая речь (женская)

Когда осуществляется алгоритм DTW, совмещение траекторий основного тона показано на рис. 3, 4. На рис. 3 показано совмещение траекторий основного тона мужской речи (целевой) с мужской речью (исходной) при использовании параметров LSF. Форма траектории основного тона деформируется, как показано в позициях цикла на рис. 3. Это соответствует совмещенному тракту, показанному на рис. 5. Видно, что при использовании параметров LSF для алгоритма DTW минимизированное значение элементов неправильно. Элемент повторен или игнорирован. Отличие форм траектории основного тона, приведенных на рис. 2 и 4, оказывается значительным. Отличие двух совмещенных трактов показано на рис. 5.

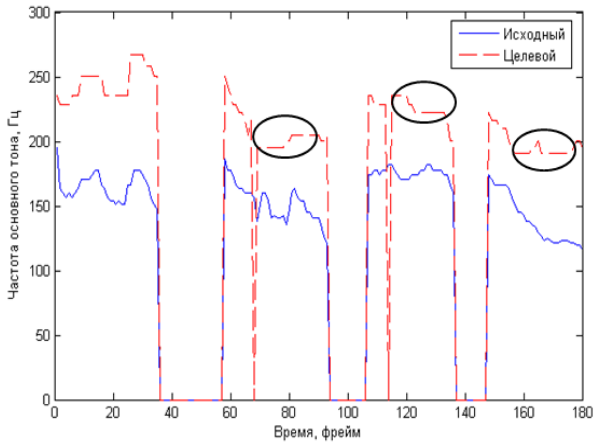

Рис. 3. Пример совмещения времени с помощью коэффициентов LSF

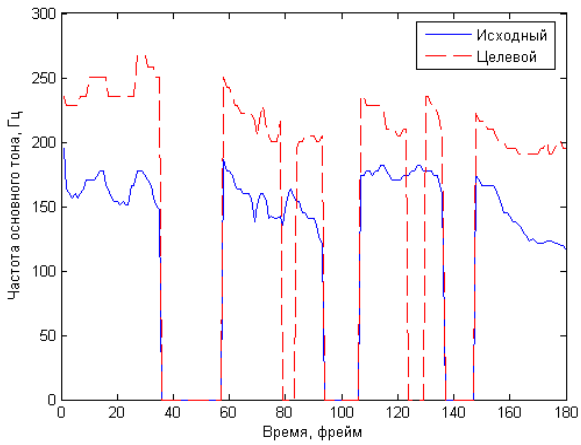

Рис. 4. Пример совмещения времени с помощью кепстральных коэффициентов

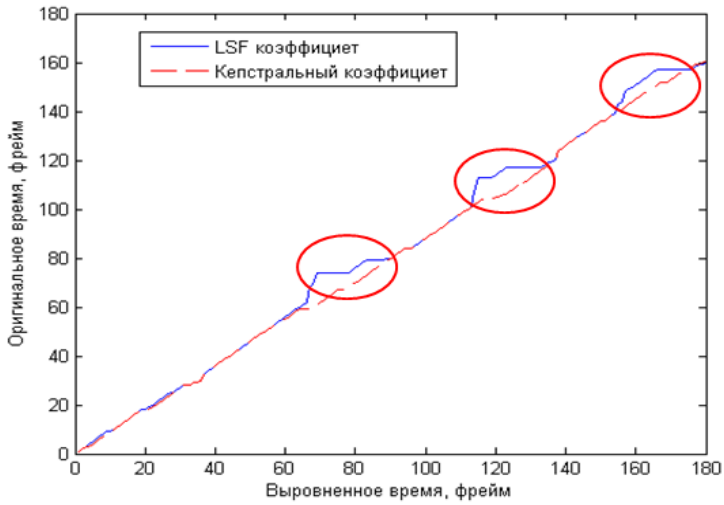

Рис. 5. Сравнение совмещенных трактов при использовании параметров LSF и кепстральных

Чтобы оценить эффективность алгоритма DTW, используя два типа параметров, используем результат совмещения времени в системах конверсии голоса. В этой системе применяется модель гауссовской смеси (GMM) [10]. Существует две основных мотивировки при использовании плотностей гауссовской смеси, как представление индивидуальности диктора [10, 11]. Первая мотивировка — интуитивное понятие о том, что отдельные компоненты плотности могут быть моделированы как множество акустических классов. Это позволяет предположить, что акустический сигнал, соответствующий голосу диктора, может характеризоваться акустическими классами. Эти классы отражают общую особенность вокального тракта вне зависимости от диктора, которая полезна для характеристики идентичности диктора. Второй мотивировкой является способность "мягкой классификации" компонентов плотности. Это является характерным свойством модели GMM. Алгоритм максимизации математического ожидания [12] используется для оценки параметров на этапе трансформации. Это и есть итеративный алгоритм, который увеличивает правдоподобие в каждой итерации с помощью успешной максимизации вспомогательной функции [13]. Основной идеей этого алгоритма является оценка нового параметра из старых. В каждой итерации алгоритма EM требуется порог (минимум ковариации), который связан с ковариацией обучаемых данных.

Обучение при использовании LSF выполняется медленнее, чем при использовании кепстральных коэффициентов. В результате эффективность спектрального преобразования обучаемых данных DTW с параметрами LSF меньше, чем такая эффективность с кепстральными параметрами. Эти результаты показаны на рис. 6 и 7. Наилучшее качество спектрального пре-

образования обучаемых данных DTW с параметрами LSF составляет 0,22. В то же время наилучшее качество спектрального преобразования обучаемых данных DTW с кепстральными параметрами составляет 0,2646.

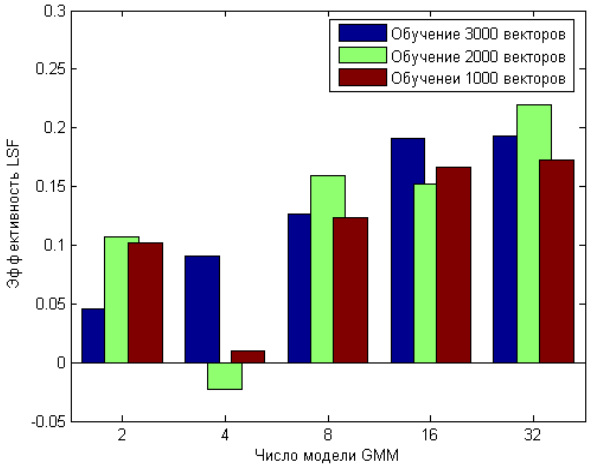

Рис. 6. Качество спектрального преобразования мужского исходного диктора (F1) в женский целевой диктор (M1) при использовании параметров LSF для совмещения времени

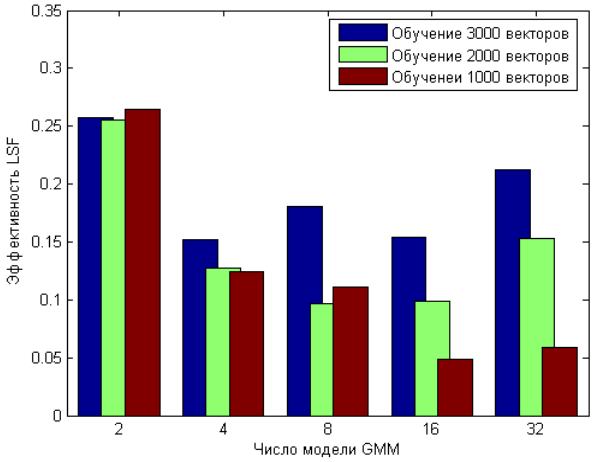

Рис. 7. Качество спектрального преобразования мужского исходного диктора (F1) в женский целевой диктор (M1) при использовании кепстральных параметров LSF для совмещения времени

Отметим, что если качество спектрального преобразования больше нуля, то достигается эффективность преобразования. Если это значение равно единице (максимальное значение), то выход системы равен исходным. Это значение может быть меньше нуля, тогда преобразование не является эффективным и выходной синтезированный сигнал системы отличается от целевого. Система VC, которая не модифицирует исходный голос, приводит к тому, что качество спектрального преобразования равно или меньше нуля, в то же время оптимальная конверсия достигается, когда качество спектрального преобразования равно единице.

Заключение

Целями данной статьи являются реализация и рассмотрение влияния выбора параметров алгоритма DTW. Алгоритм DWT вычисляет нелинейное отображение фреймов речевого сигнала в другой фрейм при помощи минимизации расстояния между ними. Это расстояние вычисляется с использованием параметров LFS и кепсральных параметров. Хотя параметры LSF имеют преимущество при кодировании речи и спектральном преобразовании, но все-таки существует недостаток при совмещении времени с использованием алгоритма DWT. На практике, спектральные параметры дают лучший результат. Обучение позволяет получать данные с преодоленной вырожденностью ковариантной матрицы и значительным улучшением конверсии голоса.

IMPLEMENTATION AND PARAMETER SELECTION OF DYNAMIC TIME WARPING ALGORITHM FOR VOICE CONVERSION

THAI TRUNG KIEN

Abstract

The Dynamic Time Warping (DTW) algorithm is widely used in speech recognition. The purpose of DTW is to produce a warping function that minimizes the total distance between the respective points (frames) of the speech signals. In this paper DTW algorithm is implemented and discussed in voice conversion area. The DTW is used to align speech frames of two sentences of two speakers. The DTW is pre-process of training phase, which will affect directly accuracy of conversion function. Selecting parameters for finding alignment path are given, in those line spectral frequencies (LSFs), cepstral coefficients are used for algorithm implementation. We see that, cepstral coefficients give the best result, which avoid singular matrix and over-fitting of training phase. The result of voice conversion system is significantly improved.

Acknowledgment

This work is made possible by the advice, experience, and support of Computer Engineering Department of Belarusian State University of Informatics and Radio Electronics. I would like to thank Prof A. A. Petrovsky for his advices and supporting.

Литература

1. *Sadaoki Furui //* Digital Speech Processing, Synthesis, and Recognition. Marcel Dekker, Inc. New York. 2001.
2. *Campbell J.P.* // Speaker Recognition: A tutorial. Proceedings of the IEEE, 85(9), September 1997. P. 1437–1462.
3. *Mizuno H.; Abe M.* // Acoustics, Speech, and Signal Processing, ICASSP-94. IEEE International Conference on. 1994. Vol 1. P. I/469–I/472.
4. *Arslan L.M.* // Speech Communication Journal. 1999. Vol. 28. P. 211–226.
5. *Chu S., Keogh E., Hart D., Pazzani M.* // Proc. of the 2nd SIAM international conference on data mining (SDM-02). 2002. P. 1–18.
6. *Paliwal K.K., Kleijn W.B.* // Quantization of LPC parameters in Speech Coding and Synthesis, Amsterdam Elsevier. 1995. P. 443–466.
7. *Fang Zheng, Zhanjiang Song, Ling Li et al.* // Int. Conf. on spoken language Processing (ICSLP-98). 1998. Vol. 3 P. 1123-1126.
8. *Wonho Yang* // Enhanced modified bark spectral distortion (EMBSD): an objective speech quality measure based on audible distortion and cognition model. Dissertation of the degree doctor of philosophy. Temple University. 1999.
9. *Рылов А.С.* // Анализ речи в распознающих системах. Минск, 2003.
10. *Stylianou Y., Cappe O.* // Acoustics, Speech and Signal Processing: proceedings of IEEE int. conf. (ICASSP - 98). Seattle, Washington, USA, May 12-15, 1998. Seattle, 1998. P. 281–284.
11. *Duxans, H.* // Proc. of EUROSPEECH Int. Conf. Geneva, Switzerland, September 1–4, 2003. Geneva, 2003. P. 861–864.
12. *Jeff A. Bilmes.* // Technical Reports International Computer Science Institute Berkeley, April 1998. Mode of access: http://www.ee.columbia.edu/~sfchang/course/svia/ee6882-paper-list.htm. Date of access: 15.05.2008.
13. *Dempster A.P., Laird N.M., Rubin D.B.* // J. of the Royal Statistical Society. 1977. Series B, Vol 34. P. 1–38.