УДК 004.934.2
doi:10.15217/issn1684-8853.2016.3.8

ПРИМЕНЕНИЕ ЧАСТОТНОГО МАСКИРОВАНИЯ ПРИ MFCC-ПАРАМЕТРИЗАЦИИ РЕЧИ НА ФОНЕ ШУМОВ

К. К. Томчук[a, 1], *старший преподаватель*
[a]*Санкт-Петербургский государственный университет аэрокосмического приборостроения, Санкт-Петербург, РФ*

Цель: *при параметризации речевых сигналов широко применяются мел-частотные кепстральные коэффициенты (MFCC), однако эффективность их использования резко падает при появлении в сигнале шумовой составляющей. Ставится задача модификации традиционного алгоритма вычисления MFCC-коэффициентов, осуществляемой путем введения дополнительных преобразований сигнала, учитывающих механизмы речеобразования и речевосприятия.* ***Результаты:*** *предложено использовать психоакустическую модель, позволяющую учитывать в расчете MFCC-коэффициентов эффект частотного маскирования при восприятии звуков человеком. Дополнительно, учитывая механизм образования в спектре речевого сигнала формантных областей, предложено воздействовать на спектральные отсчеты, соответствующие кратным гармоникам основного тона. Модифицированный алгоритм исследован на базе системы распознавания одиночных слов, адаптированной под параметризацию речевого сигнала только MFCC-коэффициентами. Показан положительный эффект от использования в алгоритме параметризации предложенных дополнительных преобразований речевого сигнала.* ***Практическая значимость:*** *представленный в работе подход к вычислению MFCC-коэффициентов сегмента речевого сигнала позволяет повысить эффективность их применения при наличии фоновых шумов в широком круге речевых приложений.*

Ключевые слова *— параметризация речевого сигнала, MFCC-коэффициенты, психоакустическая модель, речевая система.*

Введение

Одним из распространенных способов параметризации речевого сигнала (РС) является использование вектора мел-частотных кепстральных коэффициентов (MFCC-коэффициентов). Процесс вычисления данных коэффициентов учитывает ряд особенностей слухового анализатора человека, что позволяет получать хорошие результаты при их применении в речевых приложениях. Основным недостатком MFCC-коэффициентов является низкая устойчивость к шумам, что приводит к резкому ухудшению показателей соответствующих речевых приложений. Ослабление данной зависимости является актуальной задачей, которой посвящено большое количество исследований [1].

Ряд алгоритмов, направленных на увеличение эффективности MFCC-параметризации при низких отношениях сигнал/шум (ОСШ), связан непосредственно с подавлением в РС шумовой составляющей спектра, что приводит к хорошим результатам при стационарных шумах [2–4]. Для оценки спектра шума в данные алгоритмы могут дополнительно вводиться механизмы определения речевой активности [5].

Поскольку стандартный алгоритм MFCC-параметризации учитывает ряд особенностей слухового восприятия, многие исследования направлены на поиск оптимальных представлений данных особенностей.

Так, для ослабления дикторозависимости получаемых коэффициентов при их применении в системах распознавания речи может быть выполнена нормализация спектра по длине голосового тракта [6]. Для вычисления оценки спектральной плотности мощности в традиционном алгоритме используется преобразование Фурье со взвешиванием окном Хэмминга. Получаемой таким образом оценке характерна высокая дисперсия, поэтому на данном шаге в ряде работ предлагается использовать иные подходы для получения сглаженного спектра: многооконное оценивание, например на основе последовательностей Слепиана или синусоидальных окон [7, 8]; вычисление спектра с минимальной дисперсией при неискаженном отклике (MVDR-спектр) [9, 10].

В алгоритме MFCC-параметризации следующими шагами после вычисления оценки спектральной плотности мощности является ее преобразование в мел-частотную шкалу и взвешивание треугольными окнами, что также продиктовано особенностью слухового аппарата воспринимать звук неравномерными по частоте критическими полосами. В работе [11] рассмотрена замена треугольных взвешивающих окон на окна вида gammatone, которые моделируют особенности звуковосприятия в улитке человеческого уха. Данные окна также используются в ряде MFCC-подобных признаков [3, 12–14], отличающихся, помимо оконного взвешивания, представлением нелинейности частотного восприятия (например, исполь-

[1] Научный руководитель — кандидат технических наук, старший научный сотрудник, доцент кафедры радиотехнических систем Санкт-Петербургского государственного университета аэрокосмического приборостроения *Ю. А. Корнеев*.

зованием ERB-частотной шкалы) и реализацией операции нелинейного преобразования мощности.

Операция нелинейного преобразования мощности направлена на моделирование зависимости между восприятием громкости человеком и реальной интенсивностью соответствующего звука. В модификациях алгоритма MFCC-параметризации функция логарифмирования может быть заменена на степенную функцию, как более устойчивую к шумам, с показателями степени 1/3 [15], 1/10 [3, 12] и др. [10, 14].

Наконец, в ряде работ рассмотрена возможность внедрения механизма слуховой маскировки, которая в явном виде в традиционном MFCC-алгоритме отсутствует. Эффект слуховой маскировки проявляется в изменении чувствительности слуха на частотах и в моментах времени, близких к маскирующему сигналу (маскеру). В первом случае говорят о частотном, одновременном маскировании, во втором — о временном маскировании. Учет механизма временного маскирования приводит к необходимости при вычислении MFCC-коэффициентов для текущего временного окна рассматривать также сигнал в смежных временных окнах. Так, в работах [16, 17] показано улучшение характеристик системы распознавания речи при применении к спектрально-временному представлению сигнала двумерного фильтра Габора, имитирующего механизм возбуждения слуховых нейронов. В работе [18] также рассматривается вопрос внедрения в MFCC-параметризацию психоакустических моделей частотного и временного маскирования, однако в данном случае это две независимые операции. Здесь и в [19] для реализации частотного маскирования применяется модель латерального торможения, реализуемая фильтрацией спектра мощности РС.

В данной работе в качестве альтернативы применяемым в модифицированных алгоритмах MFCC-параметризации моделям частотного маскирования рассмотрен подход, широко используемый в системах сжатия аудиосигналов и заключающийся в вычислении глобального порога маскирования. Для сравнения с полученными при данном подходе показателями системы распознавания речи приведены результаты, полученные реализацией упомянутого выше механизма латерального торможения.

Также нами предложена гипотеза о возможности воздействия на спектр сигнала на частотах высших гармоник основного тона в целях улучшения MFCC-параметризации. Появление данной гипотезы связано с тем, что по длительности большую часть РС русской речи составляют вокализованные звуки, и в то же время важнейшую роль в восприятии звука играют резонансные частоты речевого аппарата (форманты), которые в свою очередь влияют на амплитудную модуляцию гармоник основного тона.

Слуховая маскировка и гармоники основного тона

В текущей работе рассматривается психоакустическая модель (ПАМ) вычисления глобального порога маскирования как альтернатива используемой в этих целях ПАМ, основанной на механизме латерального торможения. В качестве первой ПАМ используется алгоритм частотного маскирования, применяемый в стандарте сжатия ISO/IEC MPEG-1 Layer 1 [20]. Его работа детально описана в статьях [21, 22] и состоит из следующих основных шагов, выполняемых после вычисления оценки спектральной плотности мощности РС в текущем временном окне анализа:

— выделения тональных маскеров;

— разделения спектра на критические полосы;

— выделения шумового маскера для каждой критической полосы;

— прореживания: в каждой критической полосе остается не более одного маскера (тонального или шумового);

— расчета индивидуальных порогов маскирования по каждому маскеру;

— наконец, на основе индивидуальных порогов маскирования — расчета для каждой гармоники спектра глобального порога маскирования, определяющего ее слышимость в текущем временном окне.

Для сравнения эффективности ПАМ вычисления глобального порога маскирования и ПАМ, основанной на механизме латерального торможения, в работе также рассматривается подалгоритм LI (Lateral Inhibition — латеральное торможение), входящий в состав алгоритма LTFC — модифицированного алгоритма MFCC-параметризации [18].

Второй выдвигаемой гипотезой увеличения эффективности применения MFCC-параметризации, как было сказано выше, является воздействие на спектр сигнала на частотах гармоник основного тона для получения более выраженной картины формантных частот. Для воздействия на значения спектра мощности на частотах, кратных частоте основного тона (ЧОТ), предложено следующее преобразование спектральной плотности мощности:

$$P(k)=\begin{cases} P_x(k)(1+(1-0{,}5|i|)\cdot K\cdot p_v), & k=\left[n\cdot N\cdot\dfrac{f_0}{F_S}\right]+i,\ n\in\mathbb{N},\ n<\dfrac{F_S}{2f_0},\ i=\overline{-1,1} \\ P_x(k) & k\neq\left[n\cdot N\cdot\dfrac{f_0}{F_S}\right]+i,\ n\in\mathbb{N},\ n<\dfrac{F_S}{2f_0},\ i=\overline{-1,1} \end{cases}, \quad (1)$$

где $P_x(k)$ — k-й отсчет спектральной плотности мощности РС для текущего временного окна; N — длина окна фурье-преобразования; F_S — частота дискретизации РС; f_0 — оценка ЧОТ в текущем временном окне, $0 \le p_v \le 1$ — метрика наличия вокализации РС в текущем временном окне (0 для невокализованного фрагмента, 1 для вокализованного); $K \ge 0$ — коэффициент преобразования, по результатам серии экспериментов коэффициент принят равным 1,3; квадратными скобками обозначена операция округления до ближайшего целого.

Экспериментальное исследование

Для исследования результатов модификации алгоритма вычисления MFCC-коэффициентов использована система распознавания раздельно произнесенных слов из ограниченного словаря, написанная на языке MatLab доцентом Ли-Мин Ли (Lee-Min Lee) из Da-Yeh University, Тайвань [23, 24]. В качестве материала для обучения и распознавания использована англоязычная база TIDIGITS, содержащая группу из 2072 обучающих фонограмм и 2486 тестовых фонограмм. Каждая фонограмма базы содержит одно слово из словаря, который включает в себя 11 слов: цифры от нуля до девяти (ноль при этом произносится в двух вариантах: «oh» и «zero»). При составлении базы использованы голоса 94 мужчин и 114 женщин, причем дикторы обучающей и тестовой групп не пересекаются.

Исходный код принятой системы распознавания изменен таким образом, что при параметризации фонограмм используются только MFCC-коэффициенты.

Для оценки значения ЧОТ фрагмента сигнала при исследовании гипотезы использован алгоритм PEFAC [25], реализованный в тулбоксе VoiceBox вычислительной среды MatLab. Указанный алгоритм не только определяет оценку ЧОТ фрагмента РС, но также возвращает метрику p_v принадлежности фрагмента к вокализованным звукам. Тем не менее для решения данной задачи могут быть опробованы и иные алгоритмы оценки текущей ЧОТ, удовлетворяющие приведенному условию.

Для моделирования шумовой обстановки в тестовые фонограммы добавляются шумы различной природы: уличный шум, шум в поезде, шум в автомобиле, шум толпы (множественные фоновые голоса) — для ОСШ от 20 до 0 дБ с шагом 5 дБ. Обучение системы распознавания производится на исходных чистых обучающих фонограммах.

В табл. 1–4 включены результаты проведенного экспериментального исследования по эффективности распознавания при использовании модификаций алгоритма MFCC-параметризации. Представленные в таблицах значения являются усредненными по перечисленным выше типам шумов. Использованы следующие обозначения алгоритмов: MFCC(13) — традиционный алгоритм вычисления тринадцати MFCC-коэффициентов; LI — внедрение алгоритма ла-

■ ***Таблица 1.*** Частота распознавания (%) при различных ОСШ

Алгоритм	Чистый	20 дБ	15 дБ	10 дБ	5 дБ	0 дБ	0–20 дБ (среднее)
MFCC(13)	90,7	75,7	68,4	58,4	44,6	31,2	55,7
LI	89,6	75,0	68,4	59,4	46,6	33,1	56,5
MPEG1	84,5	75,8	71,9	65,1	55,0	40,3	61,6
OT	90,4	75,8	68,9	59,0	46,0	32,5	56,4
LI+OT	89,1	75,1	68,3	59,7	47,3	34,0	56,9
MPEG1+OT	84,6	75,7	71,9	65,1	55,3	40,8	61,8

■ ***Таблица 2.*** Относительные улучшения (%) по сравнению с алгоритмом MFCC(13)

Алгоритм	Чистый	20 дБ	15 дБ	10 дБ	5 дБ	0 дБ	0–20 дБ (среднее)
LI	–12,6	–2,7	0,1	2,3	3,6	2,6	1,2
MPEG1	–67,8	0,3	11,2	16,0	18,9	13,2	11,9
OT	–3,9	0,3	1,8	1,3	2,6	1,9	1,6
LI+OT	–17,8	–2,4	–0,2	2,9	4,9	4,0	1,8
MPEG1+OT	–66,5	0,1	11,2	16,1	19,4	13,8	12,1

■ ***Таблица 3.*** Частота распознавания (%) при оценке ЧОТ на чистой фонограмме

Алгоритм	Чистый	20 дБ	15 дБ	10 дБ	5 дБ	0 дБ	0–20 дБ (среднее)
OTi	90,4	76,2	69,6	59,8	46,9	33,5	57,2
LI+OTi	89,1	75,3	68,7	60,2	47,9	34,8	57,4
MPEG1+OTi	84,6	75,8	72,1	65,4	55,5	41,4	62,0

■ ***Таблица 4.*** Относительные улучшения (%) при оценке ЧОТ на чистой фонограмме

Алгоритм	Чистый	20 дБ	15 дБ	10 дБ	5 дБ	0 дБ	0–20 (среднее)
OTi	–3,9	1,9	4,0	3,4	4,3	3,3	3,4
LI+OTi	–17,8	–1,8	1,1	4,3	6,0	5,1	3,0
MPEG1+OTi	–66,5	0,2	11,7	16,6	19,8	14,8	12,6

терального торможения [18]; MPEG1 — внедрение ПАМ, представленной в стандарте ISO/IEC MPEG-1 Layer 1 [20]; OT — внедрение алгоритма усиления гармоник основного тона, формула (1); LI+OT и MPEG1+OT — совместное использование алгоритма OT соответственно с алгоритмами LI и MPEG1.

В табл. 1 приведены полученные частоты распознавания различными алгоритмами, в табл. 2 — значения относительного улучшения результатов распознавания по сравнению с алгоритмом MFCC (13). Относительное улучшение RI рассчитывается по формуле

$$RI = \frac{RR_A - RR_{MFCC}}{100 - RR_{MFCC}} \times 100\%, \qquad (2)$$

где RR_A — частота распознавания, полученная для рассматриваемого модифицированного алгоритма, %; RR_{MFCC} — частота распознавания, полученная алгоритмом MFCC(13) при тех же условиях, %.

Традиционный алгоритм MFCC-параметризации показывает лучшие результаты для чистого РС, данное положение подтверждается и другими исследованиями [1, 26]. Однако эффективность использования традиционного алгоритма стремительно падает при уменьшении ОСШ. В этом случае лучших результатов можно добиться, используя модифицированные алгоритмы. Как видно из табл. 1 и 2, внедрение в алгоритм ПАМ, широко применяемой в системах сжатия аудиосигналов и прописанной в стандарте ISO/IEC MPEG-1 Layer 1, позволяет достигнуть значимого улучшения работы речевого приложения в шумовом окружении.

Предложенное преобразование спектра мощности РС, заключающееся в усилении спектральных составляющих на частотах, кратных ЧОТ, также позволяет повысить результаты работы речевого приложения, и, как видно из представленных таблиц, данное преобразование при низких ОСШ может быть использовано совместно с психоакустическими модификациями LI и MPEG1: LI+OT, MPEG1+OT.

Поскольку используемый в модификации OT алгоритм оценки ЧОТ также подвержен влиянию фоновых шумов, для оценки потенциального эффекта использования модификации OT был рассмотрен идеализированный алгоритм оценки ЧОТ, обозначаемый далее OTi. В данной идеализации на вход используемого алгоритма оценки ЧОТ подается соответствующий фрагмент чистой фонограммы, а все остальные блоки попрежнему работают с фрагментами с заданным ОСШ. Полученные результаты представлены в табл. 3 и 4.

Производительность алгоритмов оценивается по суммарному времени, затрачиваемому вычислительной машиной на параметризацию всей базы обучающих и тестовых фонограмм. Полученные значения, нормированные к значению для традиционного алгоритма MFCC(13), приведены в табл. 5.

Психоакустическая модель, описываемая стандартом ISO/IEC MPEG-1 Layer 1, требует значительно больше вычислительных затрат, нежели фильтрация спектра мощности, реализующая алгоритм LI. Тем не менее реализации данной ПАМ обладают достаточным быстродействием

■ ***Таблица 5.*** Относительное время работы алгоритмов MFCC-параметризации

Алгоритм	Время	Алгоритм	Время
MFCC(13)	1,0	OT	3,8
LI	1,1	LI+OT	3,9
MPEG1	7,6	MPEG1+OT	11,4

для применения в системах реального времени [27]. Быстродействие модификации ОТ, в свою очередь, непосредственно определяется быстродействием применяемого в ней алгоритма оценки ЧОТ.

Заключение

Таким образом, применение в алгоритме вычисления MFCC-коэффициентов ПАМ стандарта ISO/IEC MPEG-1 Layer 1, а также предложенного преобразования спектральной плотности мощности на частотах кратных гармоник основного тона (1) позволяет получить относительное улучшение работы системы распознавания одиночных слов соответственно на 11,9 и 1,6 % при усреднении по шумам в диапазоне ОСШ 0–20 дБ.

Преобразование (1) при низких ОСШ может применяться совместно с рассмотренными ПАМ: на 0 дБ для ISO/IEC MPEG-1 Layer 1 получено дополнительное увеличение эффективности на 0,6 %, для механизма латерального торможения — на 1,4 %.

Недостатком предложенных модификаций является снижение быстродействия алгоритма, в результате чего процесс MFCC-параметризации требует на порядок больше вычислительных ресурсов по сравнению с традиционным вариантом.

Литература

1. **Majeed S. A., Husain H., Samad S. A., Idbeaa T. F.** Mel Frequency Cepstral Coefficients (MFCC) Feature Extraction Enhancement in the Application of Speech Recognition: a Comparison Study // Journal of Theoretical and Applied Information Technology. 2015. Vol. 79. N 1. P. 38–56.
2. **Tan L. N., Alwan A.** Feature Enhancement using Sparse Reference and Estimated Soft-Mask Exemplar-Pairs for Noisy Speech Recognition // Proc. IEEE Intern. Conf. ICASSP, Florence, Italy. 2014. P. 1710–1714. doi:10.1109/ICASSP.2014.6853890
3. **Chang S. Y., Meyer B. T., Morgan N.** Spectro-Temporal Features for Noise-Robust Speech Recognition using Power-Law Nonlinearity and Power-Bias Subtraction // Proc. IEEE Intern. Conf. ICASSP, Vancouver, Canada. 2013. P. 7063–7067. doi: 10.1109/ICASSP.2013.6639032
4. **Mandel M. I., Narayanan A.** Analysis by Synthesis Feature Estimation for Robust Automatic Speech Recognition using Spectral Masks // Proc. IEEE Intern. Conf. ICASSP, Florence, Italy. 2014. P. 2528–2532. doi:10.1109/ICASSP.2014.6854052
5. **Alam J., Kenny P., Dumouchel P., O'Shaughnessy D.** Noise Spectrum Estimation using Gaussian Mixture Model-based Speech Presence Probability for Robust Speech Recognition // Proc. 15th Intern. Conf. INTERSPEECH, Singapore. 2014. P. 2759–2763.
6. **Arsikere H., Alwan A.** Frequency Warping using Subglottal Resonances: Complementarity with VTLN and Robustness to Additive Noise // Proc. IEEE Intern. Conf. ICASSP, Florence, Italy. 2014. P. 6299–6303. doi:10.1109/ICASSP.2014.6854817
7. **Alam J., Kenny P., Stafylakis T.** Combining Amplitude and Phase-based Features for Speaker Verification with Short Duration Utterances // Proc. 16th Intern. Conf. INTERSPEECH, Dresden, Germany. 2015. P. 249–253.
8. **Attabi Y., Alam J., Dumouchel P., Kenny P.** Multiple Windowed Spectral Features for Emotion Recognition // Proc. IEEE Intern. Conf. ICASSP, Vancouver, Canada. 2013. P. 7527–7531. doi:10.1109/ICASSP.2013.6639126
9. **Vaz C., Tsiartas A., Narayanan S.** Energy-Constrained Minimum Variance Response Filter for Robust Vowel Spectral Estimation // Proc. IEEE Intern. Conf. ICASSP, Florence, Italy. 2014. P. 6275–6279. doi:10.1109/ICASSP.2014.6854811
10. **Alam J., Kenny P., O'Shaughnessy D.** Regularized Minimum Variance Distortionless Response-Based Cepstral Features for Robust Continuous Speech Recognition // Speech Communication. 2015. Vol. 73. P. 28–46.
11. **Slaney M., Seltzer M. L.** The Influence of Pitch and Noise on the Discriminability of Filterbank Features // Proc. 15th Intern. Conf. INTERSPEECH, Singapore. 2014. P. 2263–2267.
12. **Chang S. Y., Wegmann S.** On the Importance of Modeling and Robustness for Deep Neural Network Feature // Proc. IEEE Intern. Conf. ICASSP, South Brisbane, Australia. 2015. P. 4530–4534. doi:10.1109/ICASSP.2015.7178828
13. **Plchot O., Matsoukas S., Matejka P., Dehak N.** Developing a Speaker Identification System for the DARPA RATS Project // Proc. IEEE Intern. Conf. ICASSP, Vancouver, Canada. 2013. P. 6768–6772. doi:10.1109/ICASSP.2013.6638972
14. **Mitra V., McLaren M., Franco H., Graciarena M.** Modulation Features for Noise Robust Speaker Identification // Proc. 14th Intern. Conf. INTERSPEECH, Lyon, France. 2013. P. 3703–3707.
15. **Zhao X., Wang D.** Analyzing Noise Robustness of MFCC and GFCC Features in Speaker Identification // Proc. IEEE Intern. Conf. ICASSP, Vancouver, Canada. 2013. P. 7204–7208. doi: 10.1109/ICASSP.2013.6639061
16. **Meyer B. T., Spille C., Kollmeier B., Morgan N.** Hooking up Spectro-Temporal Filters with Auditory-Inspired Representations for Robust Automatic Speech Recognition // Proc. 13th Intern. Conf. INTERSPEECH, Portland, USA. 2012. P. 1259–1262.
17. **Kollmeier B., Schaedler M. R., Meyer A. F.** Do We Need STRFs for Cocktail Parties? On the Rel-

evance of Physiologically Motivated Features for Human Speech Perception Derived from Automatic Speech Recognition // Advances in Experimental Medicine and Biology. 2013. Vol. 787. P. 333–341. doi:10.1007/978-1-4614-1590-9_37
18. **Dai P., Soon Y.** An Improved Model of masking effects for Robust Speech Recognition System // Speech Communication. 2013. Vol. 55. P. 387–396. doi:10.1016/j.specom.2012.12.005
19. **Xugang L., Gang L., Lipo W.** Lateral Inhibition Mechanism in Computational Auditory Model and its Application in Robust Speech Recognition // Neural Networks for Signal Processing X, 2000: Proc. of the 2000 IEEE Signal Processing Society Workshop. 2000. Vol. 2. P. 785–794. doi:10.1109/nnsp.2000.890158
20. ISO/IEC International Standard 11172-3. Information Technology — Coding of Moving Pictures and Associated Audio for Digital Storage Media at up to about 1.5 Mbits/s. Part 3: Audio. — Geneva, 1993.
21. **Premananda B. S., Uma B. V.** Incorporating Auditory Masking Properties for Speech Enhancement in Presence of Near-end Noise // Intern. Journal of Computer Applications. 2014. Vol. 106. N 15. P. 1–6.
22. **Painter T., Spanias A.** Perceptual Coding of Digital Audio // Proc. of the IEEE. 2000. Vol. 88. N 4. P. 451–513.
23. **Lee L. M.** HMM Speech Recognition in Matlab. http://sourceforge.net/projects/hmm-asr-matlab/ (дата обращения: 20.09.2015).
24. **Lee L. M.** Duration High-Order Hidden Markov Models and Training Algorithms for Speech Recognition // Journal of Information Science and Engineering. 2015. Vol. 31. N 3. P. 799–820.
25. **Gonzalez S., Brookes M.** PEFAC — a Pitch Estimation Algorithm Robust to High Levels of Noise // IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP). 2014. Vol. 22. N 2. P. 518–530. doi:10.1109/TASLP.2013.2295918
26. **Dai P., Soon Y.** A Temporal Frequency Warped (TFW) 2D Psychoacoustic Filter for Robust Speech Recognition System // Speech Communication. 2011. Vol. 53. P. 229–241. doi:10.1016/j.specom.2011.10.004
27. **Noll P.** MPEG Digital Audio Coding Standards // The Digital Signal Processing Handbook/Ed. by V. K. Madisetti and D. B. Williams. — IEEE Press/CRC Press, 1998. P. 40-1–40-28.

UDC 004.934.2
doi:10.15217/issn1684-8853.2016.3.8

Frequency Masking in Speech MFCC-Parameterization in Presence of Noise

Tomchuk K. K.[a], Senior Lecturer, wake@inbox.ru
[a]Saint-Petersburg State University of Airspace Instrumentation, 67, B. Morskaia St., 190000, Saint-Petersburg, Russian Federation

Introduction: MFCCs are widely used in speech signals parameterization, however their effectiveness significantly decreases when the signal contains a noise term. The work proposes and studies a modification of the traditional MFCC calculation algorithm by introducing additional signal transformations based on the mechanisms of speech production and perception. **Results:** We propose a psychoacoustic model which allows you to take into account the frequency-masking effect in human auditory perception. In addition, considering how formant areas are formed in the voice spectrum, we propose to influence the spectral counts corresponding to multiple harmonics of the fundamental tone. The modified algorithm was tested in a single-word recognition system adapted for speech signal parameterization with only MFCCs. We demonstrated a positive effect of using the proposed additional speech signal transformations in the parameterization algorithm. **Practical relevance:** The proposed approach to MFCC calculation for the speech signal segment allows you to improve MFCC usage effectiveness in a variety of speech applications.
Keywords — Speech Signal Parameterization, MFCC, Psychoacoustic Model, Speech System.

References

1. Majeed S. A., Husain H., Samad S. A., Idbeaa T. F. Mel Frequency Cepstral Coefficients (MFCC) Feature Extraction Enhancement in the Application of Speech Recognition: a Comparison Study. *Journal of Theoretical and Applied Information Technology*, 2015, vol. 79, no. 1, pp. 38–56.
2. Tan L. N., Alwan A. Feature Enhancement using Sparse Reference and Estimated Soft-Mask Exemplar-Pairs for Noisy Speech Recognition. *Proc. IEEE Intern. Conf. ICASSP*, Florence, Italy, 2014, pp. 1710–1714. doi:10.1109/ICASSP.2014.6853890
3. Chang S. Y., Meyer B. T., Morgan N. Spectro-Temporal Features for Noise-Robust Speech Recognition using Power-Law Nonlinearity and Power-Bias Subtraction. *Proc. IEEE Intern. Conf. ICASSP*, Vancouver, Canada, 2013, pp. 7063–7067. doi:10.1109/ICASSP.2013.6639032
4. Mandel M. I., Narayanan A. Analysis by Synthesis Feature Estimation for Robust Automatic Speech Recognition using Spectral Masks. *Proc. IEEE Intern. Conf. ICASSP*, Florence, Italy, 2014, pp. 2528–2532. doi:10.1109/ICASSP.2014.6854052
5. Alam J., Kenny P., Dumouchel P., O'Shaughnessy D. Noise Spectrum Estimation using Gaussian Mixture Model-based Speech Presence Probability for Robust Speech Recognition. *Proc. 15th Intern. Conf. INTERSPEECH*, Singapore, 2014, pp. 2759–2763.
6. Arsikere H., Alwan A. Frequency Warping using Subglottal Resonances: Complementarity with VTLN and Robustness to Additive Noise. *Proc. IEEE Intern. Conf. ICASSP*, Florence, Italy, 2014, pp. 6299–6303. doi:10.1109/ICASSP.2014.6854817
7. Alam J., Kenny P., Stafylakis T. Combining Amplitude and Phase-based Features for Speaker Verification with Short Duration Utterances. *Proc. 16th Intern. Conf. INTERSPEECH*, Dresden, Germany, 2015, pp. 249–253.
8. Attabi Y., Alam J., Dumouchel P., Kenny P. Multiple Windowed Spectral Features for Emotion Recognition. *Proc.*

IEEE Intern. Conf. ICASSP, Vancouver, Canada, 2013, pp. 7527–7531. doi:10.1109/ICASSP.2013.6639126
9. Vaz C., Tsiartas A., Narayanan S. Energy-Constrained Minimum Variance Response Filter for Robust Vowel Spectral Estimation. *Proc. IEEE Intern. Conf. ICASSP*, Florence, Italy, 2014, pp. 6275–6279. doi:10.1109/ICASSP.2014.6854811
10. Alam J., Kenny P., O'Shaughnessy D. Regularized Minimum Variance Distortionless Response-Based Cepstral Features for Robust Continuous Speech Recognition. *Speech Communication*, 2015, vol. 73, pp. 28–46.
11. Slaney M., Seltzer M. L. The Influence of Pitch and Noise on the Discriminability of Filterbank Features. *Proc. 15th Intern. Conf. INTERSPEECH*, Singapore, 2014, pp. 2263–2267.
12. Chang S. Y., Wegmann S. On the Importance of Modeling and Robustness for Deep Neural Network Feature. *Proc. IEEE Intern. Conf. ICASSP*, South Brisbane, Australia, 2015, pp. 4530–4534. doi:10.1109/ICASSP.2015.7178828
13. Plchot O., Matsoukas S., Matejka P., Dehak N. Developing a Speaker Identification System for the DARPA RATS Project. *Proc. IEEE Intern. Conf. ICASSP*, Vancouver, Canada, 2013, pp. 6768–6772. doi:10.1109/ICASSP.2013.6638972
14. Mitra V., McLaren M., Franco H., Graciarena M. Modulation Features for Noise Robust Speaker Identification. *Proc. 14th Intern. Conf. INTERSPEECH*, Lyon, France, 2013, pp. 3703–3707.
15. Zhao X., Wang D. Analyzing Noise Robustness of MFCC and GFCC Features in Speaker Identification. *Proc. IEEE Intern. Conf. ICASSP*, Vancouver, Canada, 2013, pp. 7204–7208. doi: 10.1109/ICASSP.2013.6639061
16. Meyer B. T., Spille C., Kollmeier B., Morgan N. Hooking up Spectro-Temporal Filters with Auditory-Inspired Representations for Robust Automatic Speech Recognition. *Proc. 13th Intern. Conf. INTERSPEECH*, Portland, USA, 2012, pp. 1259–1262.
17. Kollmeier B., Schaedler M. R., Meyer A. F. Do We Need STRFs for Cocktail Parties? On the Relevance of Physiologically Motivated Features for Human Speech Perception Derived from Automatic Speech Recognition. *Advances in Experimental Medicine and Biology*, 2013, vol. 787, pp. 333–341. doi:10.1007/978-1-4614-1590-9_37
18. Dai P., Soon Y. An Improved Model of Masking Effects for Robust Speech Recognition System. *Speech Communication*, 2013, vol. 55, pp. 387–396. doi:10.1016/j.specom.2012.12.005
19. Xugang L., Gang L., Lipo W. Lateral Inhibition Mechanism in Computational Auditory Model and its Application in Robust Speech Recognition. *Neural Networks for Signal Processing X, 2000. Proc. of the 2000 IEEE Signal Processing Society Workshop*, 2000, vol. 2, pp. 785–794. doi:10.1109/nnsp.2000.890158
20. ISO/IEC International Standard 11172-3. *Information Technology — Coding of Moving Pictures and Associated Audio for Digital Storage Media at up to about 1.5 Mbits/s*. Part 3: Audio. Geneva, 1993.
21. Premananda B. S., Uma B. V. Incorporating Auditory Masking Properties for Speech Enhancement in presence of Near-end Noise. *International Journal of Computer Applications*, 2014, vol. 106, no. 15, pp. 1–6.
22. Painter T., Spanias A. Perceptual Coding of Digital Audio. *Proc. of the IEEE*, 2000, vol. 88, no. 4, pp. 451–513.
23. Lee L. M. *HMM Speech Recognition in Matlab*. Available at: http://sourceforge.net/projects/hmm-asr-matlab/ (accessed 20 September 2015).
24. Lee L. M. Duration High-Order Hidden Markov Models and Training Algorithms for Speech Recognition. *Journal of Information Science and Engineering*, 2015, vol. 31, no. 3, pp. 799–820.
25. Gonzalez S., Brookes M. PEFAC — a Pitch Estimation Algorithm Robust to High Levels of Noise. *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, 2014, vol. 22, no. 2, pp. 518–530. doi:10.1109/TASLP.2013.2295918
26. Dai P., Soon Y. A Temporal Frequency Warped (TFW) 2D Psychoacoustic Filter for Robust Speech Recognition System. *Speech Communication*, 2011, vol. 53, pp. 229–241. doi:10.1016/j.specom.2011.10.004
27. Noll P. *MPEG Digital Audio Coding Standards*. In: *The Digital Signal Processing Handbook*. Ed. by V. K. Madisetti and D. B. Williams. IEEE Press/CRC Press, 1998, pp. 40-1–40-28.