ЗАО «АВТЭКС СПб» zuk@autex.spb.ru 567-72-02

ОБРАБОТКА СПЕКТРА РЕЧЕВОГО СИГНАЛА

Колоколов А.С.

Институт проблем управления им. В.А. Трапезникова РАН, Москва,
kolokolo@ipu.rssi.ru.

Предлагается обработка спектра речевого сигнала, улучшающая распознавание речи в присутствии частотных искажений и аддитивных помех. В её основе используется линейная полосовая фильтрация логарифмического амплитудного спектра с последующим нелинейным преобразованием, моделирующая эффект латерального торможения в слуховом анализаторе.

В основе современных систем распознавания речи используется процедура сравнения текущего спектра речевого сигнала со спектральными эталонами, полученными в процессе обучения системы. В этом случае получается сравнительно хорошее распознавание для фиксированного диктора со стабильным произношением и неизменных характеристик окружающей среды. При нарушении этих условий качество распознавания резко снижается. Причинами этого являются значительная вариабельность произношения диктора, влияние шумов и частотных искажений, обусловленных акустикой помещения, характеристиками микрофона и свойствами канала связи. Перечисленные факторы приводят к значительным вариациям спектра речевого сигнала, что нарушает процесс сравнения с имеющимися спектральными эталонами, в результате чего значительно ухудшается качество распознавания. Тем не менее, восприятие речи человеком мало чувствительно к перечисленным выше факторам и существенно превосходит современные системы автоматического распознавания речи в точности распознавания. Это, по-видимому, связано с тем, что слуховая система располагает механизмами обработки речи, обеспечивающими получение адекватных информативных признаков речевого сигнала. Относительно природы таких признаков в настоящее время есть веские основания считать, что информация о речевом сигнале передаётся изменениями его кратковременного амплитудного спектра $S(f,t)$, где: f - частота, t - время. Такой вывод подтверждается результатами исследований по речеобразованию, акустике и фонетике, психофизике и физиологии слухового анализатора.

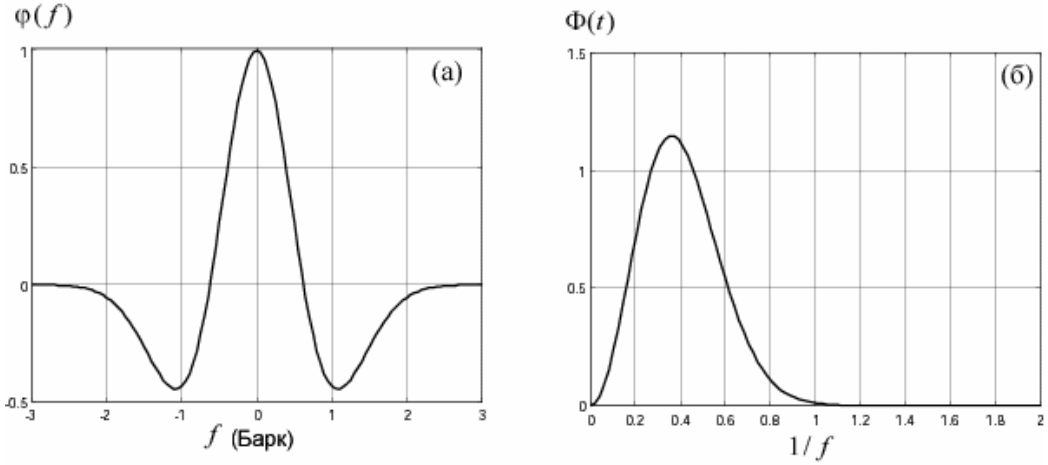

Рис.1. а – весовая функция $\varphi(f)$, б – модуль преобразования Фурье $\varphi(f)$.

В связи с этим привлекательной представляется гипотеза о том, что фонетическое качество звука связано с неоднородностями его спектра по частоте, выделяемыми благодаря эффекту латерального торможения в слуховой системе. Этот эффект обычно связывается с тем, что кратковременный амплитудный спектр $S(f,t)$, получаемый периферическим слуховым частотным анализатором, подвергается специальной локальной обработке, которую можно представить свёрткой $S(f,t)$ с весовой функцией $\varphi(f)$, описывающей характер распределения возбуждающих и тормозных связей в проводящих слуховых путях анализатора, и последующим нелинейным преобразованием. В результате формируется преобразованный спектр

(1) $S_1(f,t) = Q(S(f,t) \otimes \varphi(f))$,

где: $\otimes$ - операция свёртки, $Q(x) = x$ при $x \geq 0$, $Q(x) = 0$ при $x < 0$. Характер весовой функции $\varphi(f)$ поясняет рис.1а. Она имеет центральный положительный лепесток, представляющий распределение возбуждающих связей, и два боковых отрицательных лепестка, характеризующих распределение тормозных связей. Уравнение (1) описывает однородную нейронную сеть, обычно используемую для моделирования эффекта латерального торможения. Как известно, такая сеть при соответствующем подборе функции $\varphi(f)$ обеспечивает подчёркивание максимумов и резких перепадов в спектре $S(f,t)$. Заметим, что $\varphi(f)$,

приведённая на рис.1а, является импульсной характеристикой полосового фильтра. Для подтверждения этого на рис.1б представлен модуль преобразования Фурье $\Phi(t)$ функции $\varphi(f)$.

Результат фильтрации оказывается более интересным, если вместо амплитудного спектра $S(f,t)$ использовать логарифмический спектр $F(f,t)=\lg S(f,t)$. В этом случае уравнение (1) принимает вид $F_1(f,t)=Q(F(f,t)\otimes\varphi(f))$. Согласно линейной модели речеобразования, речевой сигнал в частотной области может быть представлен в виде произведения $S(f,t)=H(f,t)E(f,t)W(f,t)$, где $H(f,t)$ - частотная характеристика речевого тракта, $E(f,t)$ - спектр шумового или голосового источника, $W(f,t)$ - характеристика фильтра, описывающего частотные искажения речевого сигнала. После логарифмирования произведение переходит в сумму $F(f,t)=\lg S(f,t)=\lg H(f,t)+\lg E(f,t)+\lg W(f,t)$. При этом составляющие $F(f,t)$ с разной скоростью изменяются с частотой f и могут быть разделены с помощью линейной фильтрации. Составляющая $W(f,t)$, связанная с частотными искажениями речи в акустической среде или канале связи, обычно сравнительно медленно изменяется с частотой. В случае шумового источника спектр $E(f,t)$ медленно убывает с частотой со скоростью $-6\div12$ дБ/окт. Для голосового источника спектр имеет более сложный вид $E(f,t)=I(f,t)G(f,t)$, где $I(f,t)$ - спектр почти периодической последовательности δ - функций, $G(f,t)$ - спектр импульса голосового источника. Спектр $I(f,t)$ близок к последовательности гармоник с равной амплитудой и в силу этого быстро изменяется с частотой. Спектр $G(f,t)$, как и в случае шумового источника, медленно убывает с частотой со скоростью $-6\div12$ дБ/окт. Скорость изменения составляющей $H(f,t)$, определяемая резонансами речевого тракта, попадает в область средних скоростей изменения с частотой относительно всех частотных составляющих, рассмотренных выше. Поэтому, производя полосовую фильтрацию логарифмического спектра $F(f,t)$, можно в спектре $F_1(f,t)$ значительно ослабить составляющие, связанные с частотными искажениями и источником, обуславливающие вариации спектра речевого сигнала, сохранив пики, присутствующие в $H(f,t)$, связанные с резонансами речевого тракта. Тем самым оказывается возможным сделать более стабильным сравнение с эталонами при распознавании.

Процесс полосовой фильтрации $F(f,t)$ завершается выполнением нелинейного преобразования $Q(x)$. С его помощью в $F_1(f,t)$ сохраняются фрагменты, связанные с максимумами (формантами) $H(f,t)$, где отношение сигнал/шум велико. Этим обеспечивает дополнительную стабилизацию $F_1(f,t)$ при наличии фоновых широкополосных шумов со спектральной плотностью, сравнительно медленно изменяющейся с частотой.

Для иллюстрации предложенного метода на рис.2 и 3 приведены спектры речевых сегментов для звуков «ч» и «ы» в слове «четыре» для исходного и продифференцированного речевого сигнала. Логарифмические спектры сигналов получались с помощью частотного анализатора, реализованного в виде гребёнки из 35 полосовых фильтров. Для этого использовались фильтры Баттерворта четвёртого порядка с наклонами частотных характеристик 12 дБ/окт. Центральные частоты фильтров f_i ($i=1,2,...,35$) были равномерно расставлены по шкале Барков с шагом 0,57 Барк, начиная с 1,95 Барк (200 Гц по частотной шкале). Полосы пропускания фильтров были выбраны в соответствии с зависимостью критической полосы слуха от частоты равными 1,5 Барк. Для фильтрации спектра применялась весовая функция $\varphi(n)=-0{,}25\delta_k(n-2)+\delta_k(n)-\delta_k(n+2)$, где $\delta_k(n)$ - функция Кронекера, $n=...-2,-1,0,1,2,...$.

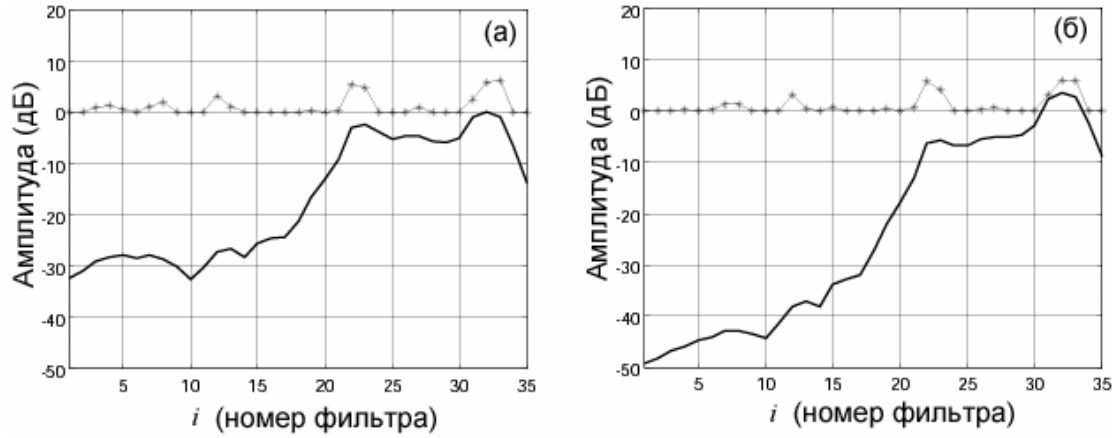

Рис.2. а – логарифмический амплитудный спектр $F(i)$ и результат его обработки $F_1(i)$ (отмечен звёздочками) для согласного «ч» в слове «четыре», б – те же зависимости для продифференцированного сигнала.

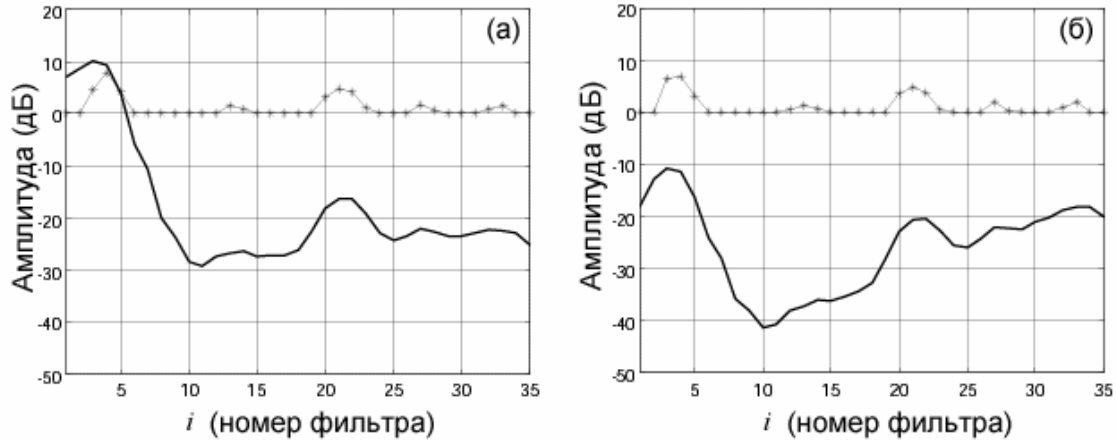

Рис.3. а – логарифмический амплитудный спектр $F(i)$ и результат его обработки $F_1(i)$ (отмечен звёздочками) для гласного «ы» в слове «четыре», б – те же зависимости для продифференцированного сигнала.

Как можно видеть из рис.2 и 3, дифференцирование речевого сигнала приводит к значительным изменениям логарифмического спектра $F(i)$, при этом обработанный спектр $F_1(i)$ остаётся практически неизменным. Подобный эффект стабилизации спектра $F_1(i)$ очевидно будет иметь место и при вариациях формы импульсов голосового источника, вызванных изменением эмоционального или психофизиологического состояния диктора.

A SPECTRUM PROCESSING OF SPEECH SIGNAL

Kolokolov A.

Institute of Control Sciences of Russian Academy of Sciences.
kolokolo@ipu.rssi.ru.

A transformation of the amplitude logarithmic speech spectrum was proposed. It takes into account the characteristics of frequency speech analysis in the auditory system. This transformation allows one to make the frequency description of the speech signal more stable in the presence of frequency distortions that slowly vary with frequency. These distortions usually arise as the result of replacement of the microphone, reverberation caused by reflections from the room walls, or changes in the form of the pulses of the voice source caused by changes in the psycho-physical state of the speaker, and so on. Operability of the method was illustrated by examples of speech signals. The results are indicative of the advisability of using it in the speech recognition systems to improve their stability to the external acoustic factors and speaker's pronunciation.