О ВОСПРИЯТИИ ИСКАЖЕНИЙ АКУСТИЧЕСКИХ СИГНАЛОВ, ОБУСЛОВЛЕННЫХ НЕЛИНЕЙНОСТЬЮ ФАЗОВОЙ ЧАСТОТНОЙ ХАРАКТЕРИСТИКИ СИСТЕМЫ

А. Н. ПРОДЕУС

Национальный технический университет Украины «КПИ», Киев

Показано, что для слуховой системы человека приемлемыми являются фазовые искажения речевых и музыкальных сигналов, если максимальная разница групповых времен задержки в области высоких и низких частот не превышает 50–70 мс.

ВВЕДЕНИЕ

Работ, посвященных исследованию восприятия слуховой системой человека искажений, обусловленных нелинейностью фазовой частотной характеристики (ФЧХ) системы, немного. В [1] отмечено, что удобной мерой нелинейности фазы $\theta(f)$ является степень неравномерности группового времени задержки $\tau(f) = -\frac{1}{2\pi}\frac{d\theta(f)}{df}$. Чувствительность слуховой системы человека к неравномерности $\tau(f)$ исследована в [2], где экспериментально показано, что искажения сигнала на слух не воспринимаются, если неравномерность $\tau(f)$ не превышает 1-3 мс. Для тренированного слуха этот порог снижается до 400 мкс. Однако упомянутые результаты были получены с использованием очень коротких, протяженностью 25 мкс, импульсов (либо небольших серий таких импульсов). В [2] отмечено, что при использовании речевых и музыкальных сигналов фазовые искажения менее заметны, однако, к сожалению, соответствующие пороговые значения неравномерности групповой задержки не определялись. Между тем, для практических целей наибольший интерес представляют именно речевые и музыкальные сигналы. Цель данной работы состояла в восполнении указанного пробела.

1. МОДЕЛЬ ВОЗНИКНОВЕНИЯ ФАЗОВЫХ ИСКАЖЕНИЙ СИГНАЛА

В [3] рассмотрена модель возникновения искажений сигнала, обусловленных нелинейностью фазовой частотной характеристики (ФЧХ) линейной системы, с использованием гребенки цифровых нерекурсивных фильтров, выходные сигналы которой суммируются (рис. 1,а). Данная модель возникновения фазовых искажений сигнала представляет значительный практический интерес, поскольку гребенки фильтров широко используются в системах записи и воспроизведения, кодирования и декодирования акустических сигналов, в линиях связи, в системах коррекции слуха [4, 5]. Хотя суммарная АЧХ гребенки может быть равномерной (Рис. 1,б), ФЧХ гребенки в общем случае является нелинейной (Рис. 1,в), даже если ФЧХ каждого из полосовых фильтров (ПФ), входящих в гребенку, является линейной. Причиной тому является различие порядка полосовых фильтров, приводящее к различным задержкам сигналов в частотных каналах (рис. 1,г).

В отличие от [1], в данной работе количество октавных фильтров гребенки увеличено с пяти до семи, что позволило охватить типичную для речевых и музыкальных сигналов полосу частот 90-11000 Гц. Основные параметры нерекурсивных фильтров,

составляющих гребенку и рассчитанных методом Ремеза, приведены в табл. 1, где f_0 - центральная частота; Δf - полоса пропускання; n - порядок фильтра.

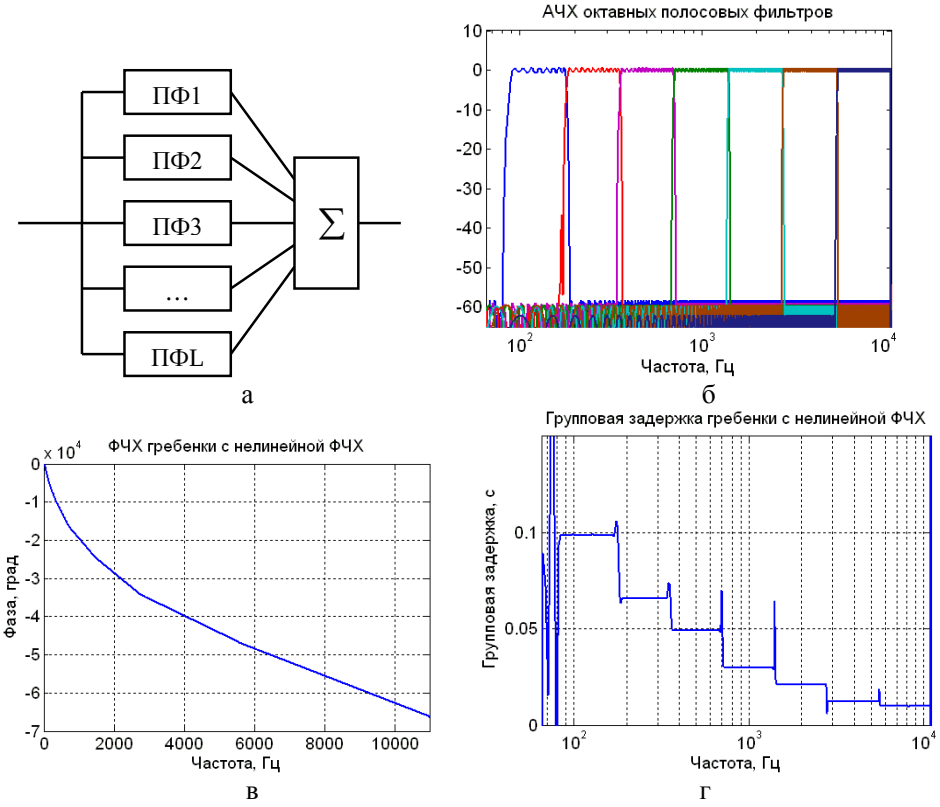


Рис. 1

Таблица 1. Параметры гребенки октавных фильтров

f_0, Гц	125	250	500	1000	2000	4000	8000
Δf, Гц	90…180	180...355	355...710	710...1400	1400...2800	2800...5600	5600…11200
n	4353	2903	2177	1320	927	545	437

На рис. 2 показана форма сигналов на входе (рис. 2а) и выходе (рис. 2б) гребенки рис. 1а для максимального различия времен задержки $\Delta\tau_{max} \approx 90$ мс, соответствующего рис. 1г. Как видим, узкополосный низкочастотный гласный звук «а» задержан на время, соизмеримое с протяженностью этого звука. Между тем, широкополосные согласные звуки «з», «д» и «ч» задержаны намного меньше. В результате согласные звуки выходного сигнала частично перекрылись с предшествующими им гласными звуками, создавая наблюдаемые визуально и воспринимаемые на слух искажения.

2. ОБЪЕКТИВНЫЕ МЕРЫ КАЧЕСТВА СИГНАЛОВ

Важной стороной рассматриваемого вопроса является выбор метода оценивания качества сигнала, а также выбор показателей качества искаженного сигнала. В работе [2]

использовались исключительно субъективные методы, очевидным и существенным недостатком которых является их высокая ресурсоемкость. Объективные (инструментальные) методы оценивания качества речевого сигнала в значительной степени свободны от указанного недостатка. Однако и здесь имеются свои трудности. Наилучшим решением было бы использование какого-либо наиболее современного показателя, такого, например, как POLQA (международный стандарт ITU-T Rec. P.863) [6], позволяющего учесть как разнообразие видов искажений сигнала, так и особенности слуховой системы человека. Однако использование POLQA в настоящее время весьма проблематично, поскольку, в силу коммерческих соображений, доступ к исходным текстам соответствующего программного обеспечения закрыт. Поэтому приходится либо использовать морально устаревший показатель PESQ (международный стандарт ITU-T Rec. P.862.2) [7], либо искать альтернативные, более простые в вычислительном плане показатели, допуская возможность их низкой эффективности.

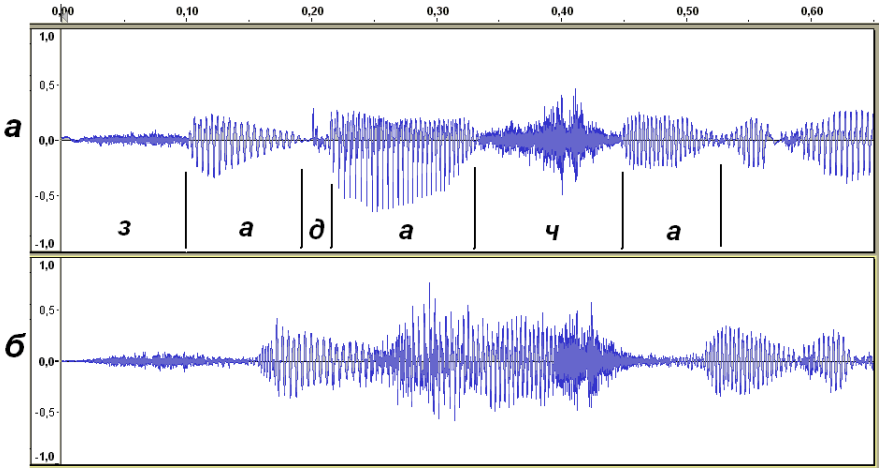


Рис. 2

Из множества известных на сегодняшний день показателей такого рода рассмотрим четыре: сегментное отношение сигнал-шум (Segmental Signal to Noise Ratio - SSNR), логарифмически-спектральные искажения (Logarithmic Spectral Distortion - LSD), барк-спектральные искажения (Bark Spectral Distortion - BSD) и перцептуальное качество речи (Perceptual Evaluation of Speech Quality - PESQ) [8, 9]. Обосновывая такой выбор, отметим, что первые два показателя – SSNR и LSD – весьма привлекательны в силу простоты вычислений, тогда как другие два показателя – BSD и PESQ – позволяют учесть, с различной степенью точности, особенности слуховой системы человека [10].

Аналитическое описание упомянутых выше показателей SSNR, LSD и BSD:

$$SSNR = \frac{1}{M}\sum_{m=1}^{M} 10\lg\left[\frac{\sum_{n=N(m-1)+1}^{Nm} x^2(n,m)}{\sum_{n=N(m-1)+1}^{Nm}[x(n,m)-y(n,m)]^2}\right], \quad (1)$$

$$LSD = \frac{2}{NM}\sum_{m=1}^{M}\sum_{r=0}^{\frac{N}{2}-1}\left|LX(r,m)-LY(r,m)\right|, \quad (2)$$

$$LX(r,m) = \max_{m,r}\{20\lg(|X(r,m)|), \delta\}, \quad \delta = \max_{m,r}\{20\lg(|X(r,m)|)\} - 50,$$

$$BSD = \frac{\sum_{m=1}^{M}\sum_{r=0}^{\frac{N}{2}-1}\left[B\{X(r,m)\} - B\{Y(r,m)\}\right]^2}{\sum_{m=1}^{M}\sum_{r=0}^{\frac{N}{2}-1}\left[B\{X(r,m)\}\right]^2}, \quad (3)$$

где $x(n,m)$ и $y(n,m)$ - n-е выборки m-го фрейма чистого сигнала $x(n)$ и искаженного сигнала $y(n)$, соответственно; $LX(r,m)$ и $LY(r,m)$ - логарифмы амплитудных спектров m-го фрейма сигналов $x(n)$ и $y(n)$, соответственно; $B\{X(r,m)\}$ и $B\{Y(r,m)\}$ - барковские спектры m-го фрейма сигналов $x(n)$ и $y(n)$, соответственно; r - номер частотной выборки; N - протяженность фреймов; M - количество фреймов. Заметим, что при практических вычислениях показателя SSNR обычно учитывают лишь те сегменты, для которых значение отношения сигнал-шум не выходит за пределы интервала [-10, +35] дБ.

Аналитическое описание алгоритма вычисления показателя PESQ весьма громоздко, поэтому в данной работе не приводится. Заметим, что следует учитывать существование двух версий PESQ – ранней (ITU-T Rec. P.862) и поздней (ITU-T Rec. P.862.2). Вторая версия, использованная в данной работе, более совершенна, поскольку позволяет анализировать речевые сигналы в широкой полосе частот (до 7 кГц).

3. РЕЗУЛЬТАТЫ ОЦЕНИВАНИЯ КАЧЕСТВА РЕЧЕВЫХ СИГНАЛОВ

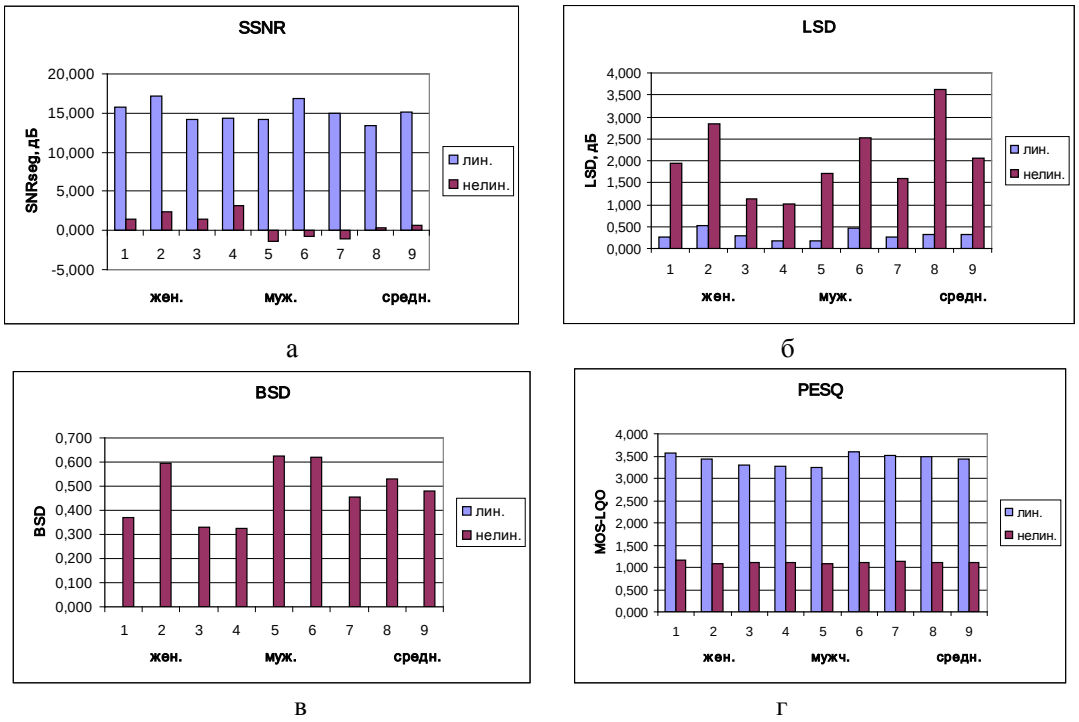
При экспериментальном оценивании, как субъективном, так и с применением указанных выше показателей качества, зависимости качества сигнала от степени нелинейности ФЧХ, использованы фрагменты, протяженностью 1 минута каждый, речевых сигналов для 4-х дикторов-женщин и 4-х дикторов-мужчин, читающих русский текст по юридической тематике. Запись сигналов произведена на кафедре акустики НТУУ «КПИ», в заглушенном помещении с временем реверберации 0,15 с, с частотой дискретизации 22050 Гц и битовой глубиной 16 бит.

Оценки объективных показателей при $\Delta\tau_{max} \approx 90$ мс для речевых сигналов, пропущенных через гребенку рис. 1а с ФЧХ, показанной на рис. 1в приведены на рис. 9, где первые четыре пары столбцов соответствуют дикторам-женщинам, вторые четыре пары – дикторам-мужчинам, последняя пара столбцов представляет средние по всем дикторам результаты. Нетрудно видеть, что все рассмотренные показатели качества адекватно отреагировали на нелинейность ФЧХ гребенки фильтров, засвидетельствовав существенное ухудшение качества речевого сигнала.

Варьирование степенью неравномерности $\Delta\tau_{max}$ позволило субъективно оценить пороговое значение чувствительности слуховой системы человека к фазовым искажениям речевого сигнала. Оказалось, что на слух первые признаки искажений в виде небольшой «сиплости» звучания речи появляются при $\Delta\tau_{max} \approx 50$ мс. Увеличение $\Delta\tau_{max}$ до 70-90 мс приводит к росту искажений речи, которые на слух воспринимаются как эффект «сиплый хорус», т.е. как одновременное чтение текста несколькими дикторами со слегка осипшими голосами.

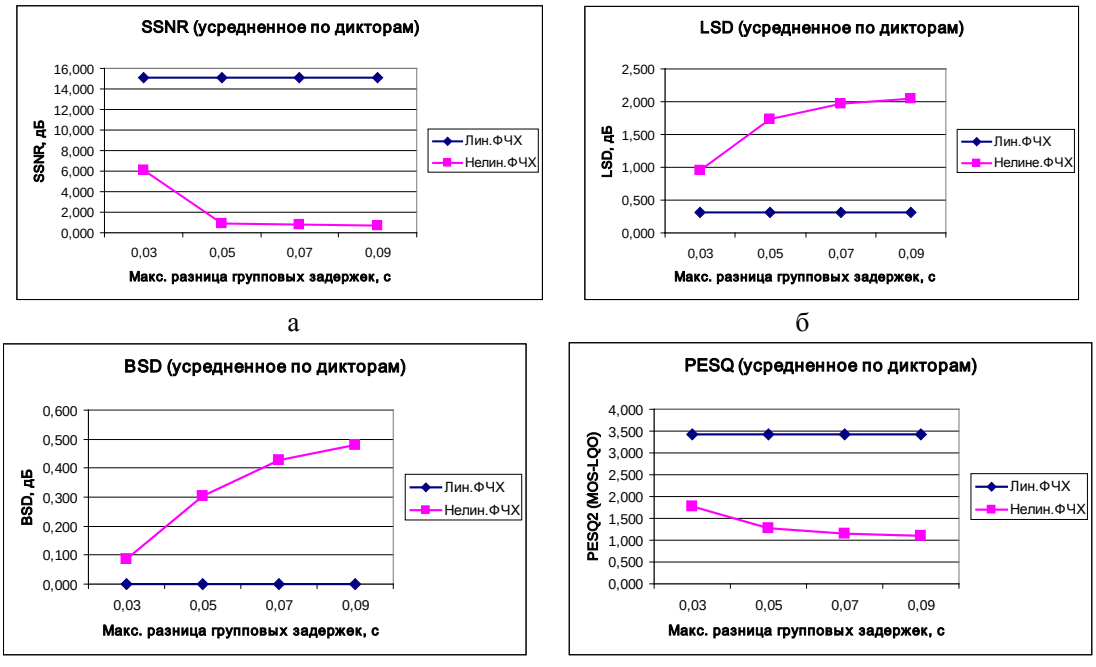
Зависимости усредненных (по дикторам) оценок объективных показателей качества от $\Delta\tau_{max}$ представлены на рис. 4.

Как следует из рис. 4, пороговому значению $\Delta\tau_{max} \approx 50$ мс соответствуют следующие пороговые значения объективных показателей: 1 дБ для SSNR; 1,8 дБ для LSD; 0,3 дБ для BSD и 1,4 MOS для PESQ.



а б в г

Рис. 3



а б в г

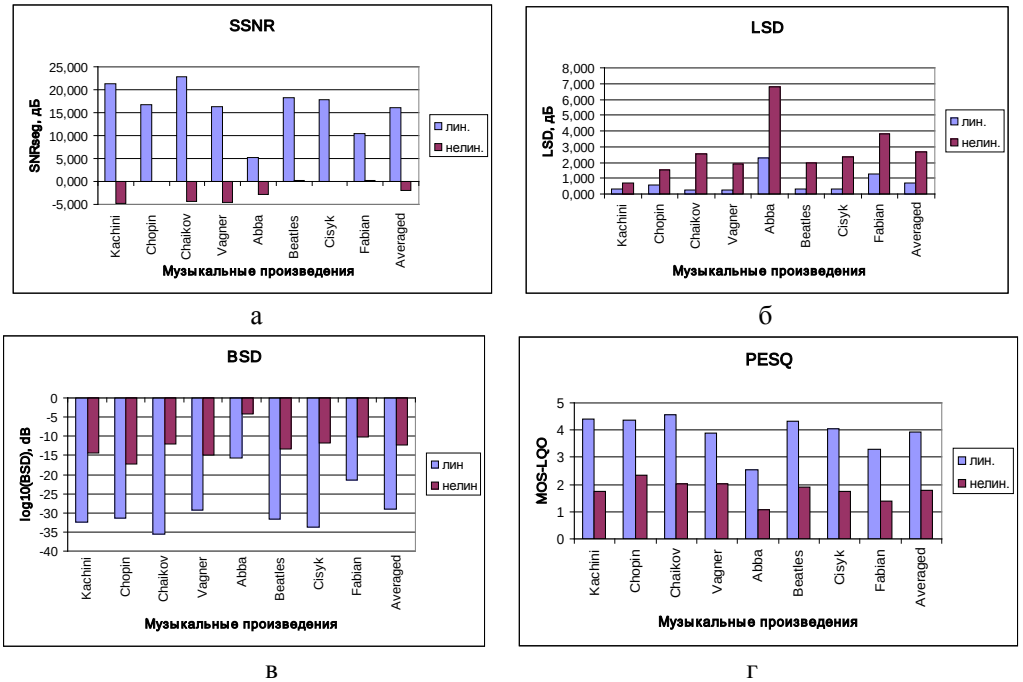
Рис. 4

4. РЕЗУЛЬТАТЫ ОЦЕНИВАНИЯ КАЧЕСТВА МУЗЫКАЛЬНЫХ СИГНАЛОВ

При экспериментальном оценивании зависимости качества музыкального сигнала от характера и степени нелинейности ФЧХ фильтра использованы фрагменты восьми музыкальных произведений протяженностью 30-45 секунд каждый. При этом половина произведений принадлежала жанру «классическая музыка» («Ave Maria» Дж. Каччини, «Этюд №4, соч. 10, Ф. Шопена», 5-я симфония П. Чайковского, увертюра «Фауст» Р. Вагнера), а половина – жанру «популярная музыка» («Mama-mia» ABBA, «Shes_Leaving_Home» Beatles, «Я піду в далекі гори» К. Цісик, «Mademoiselle Hyde» L. Fabian). Все сигналы записаны с частотой дискретизации 22050 Гц и битовой глубиной 16 бит.

Оценивание качества звучания производилось как на слух, так и с использованием перечисленных выше объективных показателей SSNR, LSD BSD и PESQ (рис. 5). Нетрудно видеть, что все рассмотренные показатели качества адекватно отреагировали на нелинейность ФЧХ гребенки фильтров, засвидетельствовав существенное ухудшение качества музыкального сигнала. Вместе с тем, приведенные графики свидетельствуют, что степень такого ухудшения неодинакова как для различных жанров, так и для отдельных произведений фиксированного жанра.

Отметим, что для музыкальных сигналов оцениваемое на слух пороговое значение $\Delta\tau_{\max}$, при котором становятся заметными фазовые искажения, несколько выше, чем для речевых сигналов, и близко 70 мс. Данный вывод согласуется с результатами объективного оценивания (рис. 6), где максимальная крутизна графиков SSNR, BSD и PESQ наблюдается для $\Delta\tau_{\max} \le 70$ мс (исключением является показатель LSD, скорость роста которого увеличивается и для $\Delta\tau_{\max} > 70$ мс).



а б

в г

Рис. 5

Наиболее заметны фазовые искажения музыки в тех произведениях, где высока роль ударных инструментов. На рис. 7 показана форма сигналов солирующих ударных для ситуации, когда максимальная разница групповых времен задержки ФЧХ в области высоких и низких частот составляет 90 мс. Кружками обведены импульсы бас-барабана, квадратом обведен импульс барабана-тома, квадратом с закругленными углами обведен импульс тарелки-крэша.

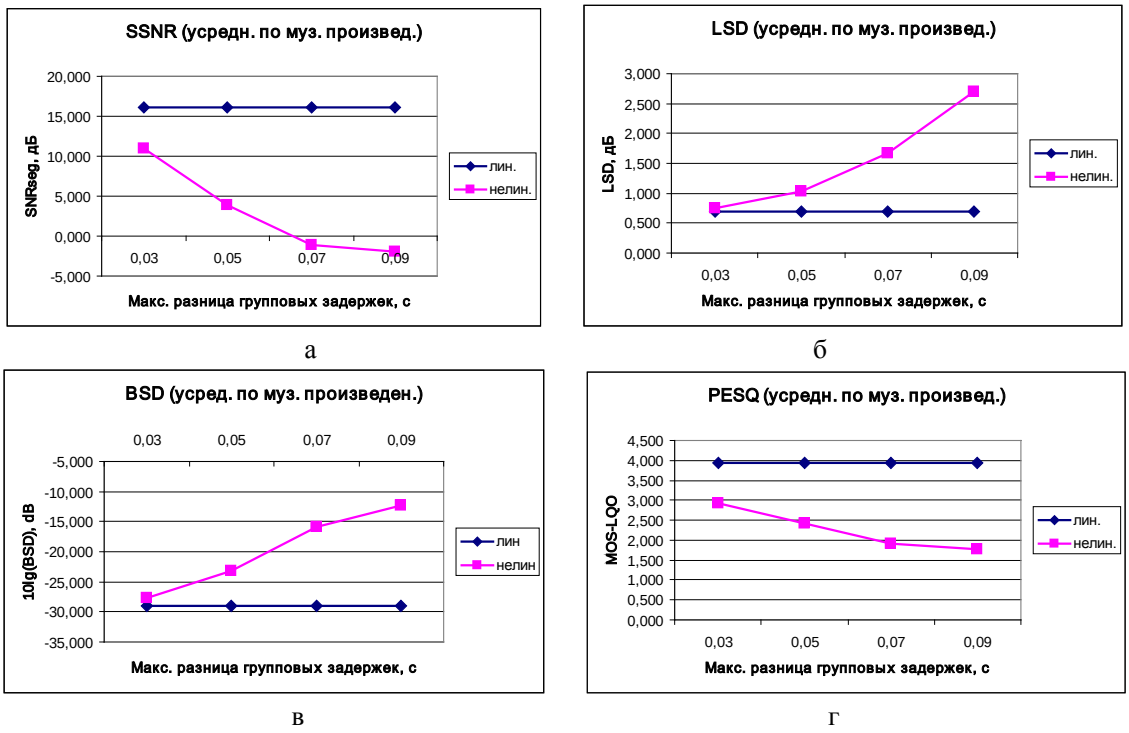


Рис. 6

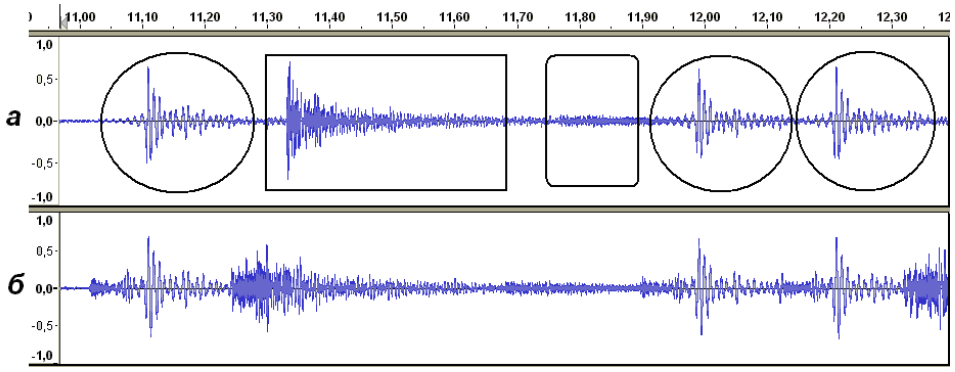


Рис. 7

ВЫВОДЫ

Экспериментальные исследования показали, что для слуховой системы человека приемлемыми являются фазовые искажения речевых сигналов, если максимальная разница групповых времен задержки в области высоких и низких частот не превышает 50 мс. Объяснить это можно тем, что при такой разнице групповых времен задержки

интерференция между смежными гласными и согласными звуками на слух практически незаметна. Пороговому значению $\Delta\tau_{max} \approx 50$ мс соответствуют следующие пороговые значения объективных показателей: 1 дБ для SSNR; 1,8 дБ для LSD; 0,3 дБ для BSD и 1,4 MOS для PESQ.

Фазовые искажения музыкальных сигналов практически не ощутимы, если максимальная разница групповых времен задержки в области высоких и низких частот не превышает 70 мс. При этом пороговому значению $\overline{\Delta\tau_{max}} \approx 70$ мс соответствуют следующие пороговые значения объективных показателей: минус 1 дБ для SSNR; 1,7 дБ для LSD; минус 15 дБ для BSD и 1,9 MOS для PESQ.

ЛИТЕРАТУРА

1. *Оппенгейм А., Шафер Р.* Цифровая обработка сигналов. – М.: Техносфера, 2006. – 858 с.
2. *Blauert J.* Group delay distortions in electroacoustical systems // J. Acoust. Soc. Am. – 1978. – **63**, № 5. – P. 1478–1483.
3. *Дидковский В. С., Дидковская М. В., Продеус А. Н.* Акустическая экспертиза каналов речевой коммуникации. – К.: Имэкс-ЛТД, 2008. – 420 с.
4. *Advances* in Digital Speech Transmission. Edited by Martin R., Heute U. and Antweiler C. – John Wiley & Sons Ltd, England, 2008. – 572 p.
5. *Communication* acoustics. Edited by Blauert J. // Berlin/Heidelberg/New York: Springer, 2005. – 385 p.
6. *Perceptual* Objective Listening Quality Assessment (POLQA) ITU-T Recommendations P.863 – January 2011.
7. *Perceptual* Evaluation of Speech Quality (PESQ) ITU-T Recommendations P.862, P.862.1, P.862.2. Version 2.0 – October 2005.
8. *Springer* Handbook of Speech Processing. Edited by Benesty J., Sondhi M. and Huang Y. Springer-Verlag Berlin, Heidelberg, 2008. – 1159 p.
9. *Naylor P., Gaubitch N.* Speech Dereverberation. – Springer, 2010. – 399 p.
10. *Bogdanova N., Prodeus A.* Objective quality evaluation of speech band-limited signals // Electronics and Communications. – 2014. – **19**, № 6(83). – P. 58–65.