АКУСТИЧЕСКИЙ ЖУРНАЛ, 1997, том 43, № 1, с. 119–120

КРАТКИЕ СООБЩЕНИЯ

УДК 534.782

КОМПАНДИРОВАНИЕ РЕЧЕВОГО СИГНАЛА ПРИ СНИЖЕНИИ ЧАСТОТЫ ЕГО ЦИФРОВОЙ ЗАПИСИ (К ПРОБЛЕМЕ ОПТИМАЛЬНОЙ ЧАСТОТЫ ДИСКРЕТИЗАЦИИ ЗВУКОВ РЕЧИ)

© 1997 г. **Л. В. Шеншев**

Психологический институт РАО
103009 Москва, ул. Моховая, 9, корп. В

Поступила в редакцию 20.11.95 г.

Объем цифровой фонограммы речи прямо пропорционален частоте дискретизации ее сигнала. Это непосредственно вытекает из самого принципа цифровой записи аналоговых сигналов, тем не менее выдвигаемый в данном сообщении тезис о возможности компандировать* речевой сигнал путем снижения частоты его цифровой записи не тривиален, поскольку подразумевает снижение частоты дискретизации значительно ниже того критического уровня (назовем его частотным порогом), начиная с которого традиционные методы порождают в декомпрессируемой речи машинный акцент и прочие акустические изъяны, нарушающие естественность ее звучания.

Среди разнообразных сфер человеческой деятельности, требующих компандирования речевого сигнала, наиболее жесткие требования к естественности звучания декомпрессируемой речи предъявляет автоматизированное обучение устной речи на иностранном языке [1], [2]. Если в военном деле, на транспорте и в промышленности еще можно мириться с машинным акцентом, то при обучении языку он недопустим. Повышены здесь требования также и к экономической стороне проблемы.

В этой связи в статье [1] были предложены два экономичных способа сжатия цифровых фонограмм, предназначенных для обучения языку. В основе одного из них лежит экстраполяция так называемых фонемных (сегментных) ядер, а в основе другого – адаптивная интерполяция (интерполяция малоинформативных отсчетов сигнала). Преимущество экстраполяционного метода – возможность сжимать звуки речи в десятки раз, а также сочетать декомпрессию речи с регулированием ее темпа. Но область его применения ограничена гласными и сонорными звуками. Интерполяционный же метод оказался универсальным, пригодным для любых речевых звуков (по крайней мере русского языка, на материале которого проводились исследования). Однако достигаемая им степень сжатия несопоставимо ниже, чем у экстраполяционного метода. В зависимости от особенностей сжимаемого звука и исходной частоты его цифровой записи она редко поднималась выше 40%.

В данном сообщении описываются результаты дополнительного исследования возможностей универсального (интерполяционного) метода. Суть исследования поясним на конкретном примере одной из серий экспериментов.

Голос, на материале которого их предстояло провести, был сначала тестирован под углом зрения присущего ему частотного порога дискретизации. Для этого частота цифровой записи произносимых этим голосом слов постепенно снижалась от 10 кГц до появления первых признаков машинного акцента или иных отклонений декомпрессируемой речи от ее исходного (не сжимавшегося) варианта. Искомый порог был зафиксирован примерно на уровне 8 кГц.

В последующих экспериментах эта величина служила точкой отсчета для дальнейшего снижения частоты дискретизации. В одном из экспериментов она была понижена до 4 кГц, т.е. вдвое. С помощью особой процедуры (на которой остановимся несколько ниже) записанная с такой частотой фонограмма подвергалась n-кратному растягиванию в сочетании с n-кратным превышением частоты воспроизведения над частотой записи.

Так, при $n = 3$ фонограмма, записанная с частотой 4 кГц, после преобразования воспроизводилась с частотой 12 кГц, а при $n = 4$ тот же самый текст, произнесенный тем же диктором и тоже записанный с частотой 4 кГц, после преобразования цифровой записи воспроизводился с частотой 16 кГц. Для сравнения тот же текст записывался с частотой 8 кГц и без всяких преобразований фонограммы воспроизводился с такой же частотой.

---

* Утвердившийся в мировой литературе термин "компандирование" (англ. compress – сжимать и expand – расширять) в принципе можно было бы заменить более понятным выражением "обратимое сжатие". Но оно воспринималось бы как тавтология, поскольку о сжатии речевых звуков принято говорить, только когда известен способ их декомпрессии, позволяющий их по мере надобности восстанавливать (аппроксимировать) с той или иной точностью.

И при $n = 3$, и при $n = 4$ декомпрессированная речь звучала не менее естественно, чем исходная, т.е. записанная в контрольной фонограмме. При $n = 3$ декомпрессированную и исходную речь еще можно было различать по тембру, но при $n = 4$ исчезло и это различие.

Преобразование (восстановление) сигнала, сжатого благодаря снижению частоты его цифровой записи до 4 кГц, протекало в два этапа (в принципе их не обязательно разносить во времени, излагаемая процедура декомпрессии может протекать "на проходе", т.е. с запоминанием минимальной промежуточной информации). На первом этапе декомпрессируемая фонограмма растягивалась благодаря тому, что ее точки (отсчеты мгновенного значения уровня сигнала) раздвигались на интервалы по $n - 1$ позиций. Тем самым внутри каждой пары соседних отсчетов формировалось по $n - 1$ "дыр" (пустот), которые затем (на втором этапе) заполнялись результатами интерполяции по формуле

$$m(i, j) = h(j) + [h(j + 1) - h(j)]i/n,$$

где

n – как и выше, показатель кратности снижения частотного порога дискретизации,

$h(j)$ – уровень j-й точки дискретизации декомпрессируемого сигнала,

i – порядковый номер очередной "дыры" между j-й и $(j + 1)$-й точками дискретизируемого сигнала,

$m(i, j)$ – уровень, присваиваемый i-й "дыре" в j-й паре соседних точек декомпрессируемого сигнала.

То, что компандируемый таким путем речевой сигнал при воспроизведении практически ничем не отличался от исходного, означает, что по крайней мере двукратное понижение частотного порога дискретизации можно полностью компенсировать интерполяцией выпавших из фонограммы отсчетов в сочетании с многократным превышением частоты воспроизведения над частотой записи.

Важно подчеркнуть, что такое компандирование не заменяет адаптивный интерполяционный метод, изложенный в статье [1], а дополняет его, расширяя его возможности. Так, в одном из экспериментов речь, записанная с частотой 4 кГц и тем самым уже сжатая вдвое, перед растягиванием была предварительно сжата на 24% методом адаптивного интерполяционного сжатия (растягиванию подверглась фонограмма, с помощью буфера декомпрессированная после такого дополнительного сжатия). Суммарное сжатие достигло, таким образом, 62%, т.е. стало без малого трехкратным.

Во всех экспериментах естественность звучания декомпрессированной речи оценивалась на основе ее сравнения с исходной, не сжимавшейся (чтобы облегчить сравнение этих ее двух вариантов, они непрерывно чередовались друг с другом в циклически воспроизводимой фонограмме). Оказалось, что естественность звучания декомпрессируемой речи не пострадала даже от комбинированного (трехкратного) сжатия, а это означает возможность по меньшей мере столь же значительного снижения частотного порога дискретизации звуков речи.

Работа выполнена при финансовой поддержке Российского гуманитарного научного фонда (код проекта 94-06-19-911-А).

СПИСОК ЛИТЕРАТУРЫ

1. *Шеншев Л.В.* Сжимаемость речевых звуков при их адаптивной дискретизации (к проблеме синтеза естественно звучащей речи) // Акуст. журн. 1995. Т. 41. № 2. С. 329–335.

2. *Шеншев Л.В.* Основы адаптивного обучения языку (семиотические аспекты развития речи с помощью автомата). М.: Наука, 1995. 113 с.