Возможности использования лингвистической информации при разработке перспективных технологий сжатия речевого сигнала

Н. В. Бобров*
Московский государственный лингвистический университет
Россия, 119034 Москва, ул. Остоженка, д. 38 стр. 1
(Статья поступила 26.06.2017; Подписана в печать 11.09.2017)

Известно, что оцифрованный речевой сигнал обладает относительно малой сжимаемостью. Компрессия речевого сигнала без потерь методами энтропийного сжатия даёт уменьшение объёма результирующего сообщения по отношению к исходному приблизительно в полтора раза. Применение дельта-компрессии (и близких к ней по сути методов, основанные на идеях А. Хаара и И. Добеши) позволяет улучшить этот показатель ещё в полтора раза. Методы сжатия речевого сигнала с потерями, дающие коэффициенты сжатия, лучшие на порядок, используют психоакустические закономерности: из сигнала удаляются компоненты, оказывающие наименьшее влияние на слуховое восприятие сигнала, например участки спектра, находящиеся «в тени» больших пиков. Необходимо заметить, что все перечисленные идеи основаны на существовании некоторых априорных знаний о речевом сигнале как источнике данных (например, о том, что ординаты соседних точек осциллограммы часто различаются на небольшую величину, или о том, что мгновенный спектр речевого сигнала, как правило, содержит небольшое число доминирующих пиков, определяющих воспринимаемое качество звука). Следуя этой же логике, можно предположить, что использование априорных знаний о том, что речевой сигнал является контейнером, заключающим в себе лингвистическую информацию, также может дать существенный выигрыш в степени его сжатия как с потерями, так и без потерь информации за счёт включения в модель источника данных сведений о закономерностях, описывающих его лингвистическую составляющую. О том, к каким результатам может привести проверка данного предположения, и пойдёт речь в предлагаемом докладе.

PACS: 43.72+q. УДК: 81.32, 81.33.
Ключевые слова: речевой сигнал, алгоритмы сжатия, кодеки, лингвистическая информация.

ВВЕДЕНИЕ

Информационная избыточность речи в её символьном (фонематическом или текстовом) представлении составляет порядка 50%, а в отдельных случаях может достигать 80% [1] и более. Этим объясняется тот факт, что речевое сообщение, переданное даже со значительным количеством ошибок, может быть корректно понято принимающей стороной. Информационная избыточность делает возможным успешное использование речевого канала коммуникации в условиях шумов и помех различного рода [2, 3] и в этом смысле является не изъяном, а сформировавшимся в ходе эволюции полезным фундаментальным свойством речи как сигнальной системы, общим для всех существующих на Земле естественных языков. Полезным оно оказывается и применительно к рассматриваемой задаче сжатия речевой информации, однако поиск способов его обращения во благо в данном случае оборачивается непростой исследовательской задачей.

1. ПРОБЛЕМА ОЦЕНКИ ИНФОРМАЦИОННОЙ ИЗБЫТОЧНОСТИ РЕЧИ

Информационная избыточность звучащей речи значительно превосходит избыточность речи в символьной записи. Чтобы убедиться в этом, достаточно рассмотреть следующий пример. Символьное сообщение «рейс SU555 прибыл по расписанию» можно закодировать 31 байтами (например, используя ASCII). Это же сообщение, произнесённое голосом (3,2 с звучания), оцифрованное при частоте АЦП 16 кГц и глубине квантования по уровню 16 бит, в несжатом виде занимает 100 килобайт. В то же время реальный объём информации, передаваемый этим сообщением, зависит от коммуникативной ситуации, в данном случае — от количества рейсов, прибытие которых ожидается в данный момент, и от множества факторов, влияющих на априорную оценку вероятности задержки этого конкретного рейса, и может составлять от 1 бита (рейс всего один, прибытие или неприбытие вовремя равновероятно) до нескольких байт (рейсов много, аэропорт вот–вот закроют по метеоусловиям).

Чтобы ответить на вопрос об информационной избыточности речи применительно к нашей задаче — сжатию передаваемого речевого сообщения — необходимо вначале уточнить, как именно следует в данном случае определять энтропию звучащего речевого сообщения. Для этого необходимо рассмотреть информационную структуру речевого сообщения и понять, какие её компоненты следует считать несущими полезную информацию, а какие — применительно к данной задаче — можно приравнять к шуму (Следует заметить, что именно шум, если он интерпретируется как подлежащие сохранению полезные данные, как правило, обладает наихудшей сжимаемостью, так как не подчи-

*E-mail: arctangent@yandex.ru

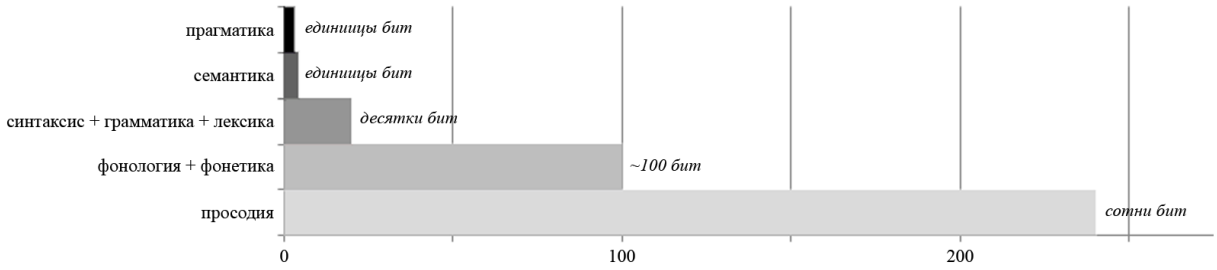

Рис. 1: Оценки энтропии различных уровней организации лингвистической и паралингвистической информации порождаемого короткого речевого сообщения

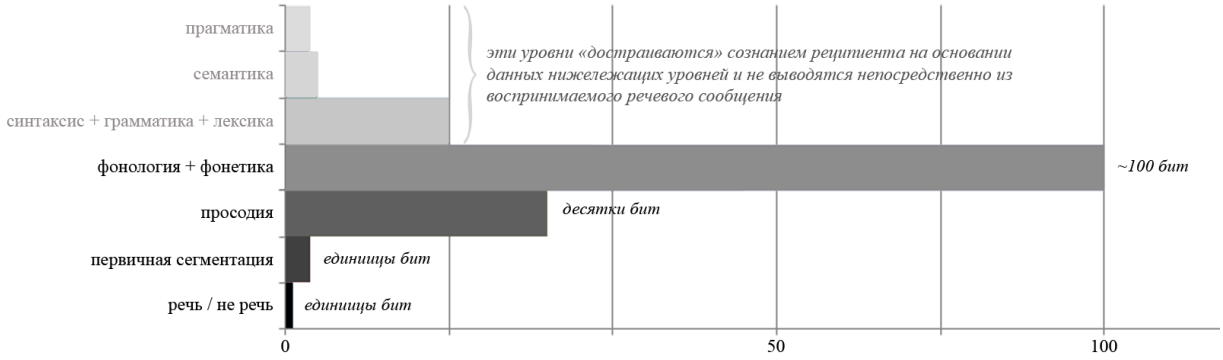

Рис. 2: Оценки энтропии различных уровней организации лингвистической и паралингвистической информации воспринимаемого короткого речевого сообщения

няется осмысленным — а стало быть, и известным — закономерностям.).

Информационная структура речевого сообщения содержит множество уровней, соответствующих уровням организации естественного языка (сверху вниз): прагматическому, семантическому, синтактико-грамматическому, лексическому, морфологическому, просодическому, фонологическому, фонетическому – и, наконец, акустическому – уровню материального носителя единиц плана выражения [4]. Оценки энтропии различных уровней организации лингвистической и паралингвистической информации речевого сообщения представлены на рис. 1–2 (Подробное изложение процедуры получения данных оценок приведено в статье [5]).

Заметим, что оценки энтропии речевого сообщения при анализе «сверху вниз» — от прагматики к фонетике и акустике — и «снизу вверх» — от акустики к прагматике — получаются разными. Это обстоятельство связано с существованием различных априорных знаний у говорящего и слушающего в момент передачи сообщения и может быть названо фундаментальным свойством асимметрии речевой коммуникации. Характерная величина суммарной энтропии короткого речевого высказывания (длительностью ~ 3 с) при оценке «сверху вниз» (то есть *со стороны говорящего*) составляет 45–50 байт, или 15–16 байт в секунду. Это больше, чем объём символьной записи, но всё равно в 2000 раз меньше, чем объём фонограммы. Объём информации, декодируемой при *восприятии* одного короткого речевого сообщения в общей сложности составляет порядка 130 бит, или 16–17 байт.

Из этого можно сделать следующие выводы. Во-первых, суммарная величина ёмкости лингвистических уровней информационной структуры речевого сообщения (включая сюда как вербальные, так и паравербальные уровни, что принципиально важно) оказывается одного порядка с характерным объёмом символьной записи высказывания. Во-вторых, «битрейт» воспринимаемой речевой информации по полученной оценке составляет около 40–45 бит (5–6 байт) в секунду, что соотносится с данными, представленными в работе [6], где, в числе прочего, по данному поводу сделано следующее замечание: «Согласно теории информации, акустический сигнал имеет объём информации порядка 300000 бит/с. Человеческий мозг в состоянии переработать примерно 50 бит/с. Из сравнения данных чисел ясно, какую работу совершает слуховой анализатор при восприятии речи в каждую секунду активного слушания, редуцируя информацию» [6, 7]. В-третьих, оценки информативности фонетического и фонологического уровней ($\sim$ 100 бит), полученные для процессов речевосприятия и речепроизводства, *совпадают*. В-четвёртых, оценки информационной ёмкости просодии для речевосприятия и речепроизводства различаются *на порядок* (в 8 раз), что в данном случае может

служить иллюстрацией трудноформализуемости паравербальной компоненты речевой коммуникации.

2. ПРИЛОЖЕНИЕ ОЦЕНОК ИНФОРМАТИВНОСТИ РЕЧЕВОГО СООБЩЕНИЯ К ЗАДАЧЕ КОМПРЕССИИ РЕЧЕВОГО СИГНАЛА

Какие из сделанных наблюдений можно вывести заключения применительно к вопросу об информационной избыточности и задаче сжатия речевого сигнала? На первый взгляд может показаться, что приведённые выше выкладки определяют некий *теоретический потолок* сжимаемости речи. Между тем, это совсем не так. Проведя все эти рассуждения, мы определили информационную ёмкость *лингвистического каркаса* речевого сообщения, используя который, теоретически возможно организовать сжатие речевого сигнала по схеме «анализ > выделение каркаса > синтез по правилам > разностное кодирование оставшейся вариативности (сначала в спектральной, а потом и во временной области)». Таким образом можно получить алгоритм сжатия, многократно превосходящий по эффективности все известные на текущий момент аналоги (даже если использовать «верхние» оценки информационной ёмкости лингвистического каркаса, речь будет идти лишь о сотнях байт на одно короткое речевое сообщение). Своеобразной «платой» за это будет включение большого объёма *статических* данных (единицы гигабайт: речь идёт, большей частью, о словаре; если оценить объём словаря в 10^6 единиц и объём описания одной единицы в 1 Кб, получится как раз 1 Гб; все остальные необходимые данные по сравнению со словарём пренебрежимо малы) в модули кодирования/декодирования речевого сигнала. Однако необходимо заметить, что такие объёмы данных представляли серьёзное препятствие на пути разработки речевых кодеков на заре их развития, но не сейчас, поэтому указанный подход к сжатию речевого сигнала может перейти из утопии в реальность. *Но это ещё не всё.* Известно, что распределение многих (вероятно, даже *всех*, но это утверждение требует дополнительной проверки) единиц, составляющих тексты (и в том числе — звучащие тексты) на естественном языке, подчиняется закону Ципфа [7]. Это означает, что выделение лингвистического каркаса речевого сигнала не только обнаруживает возможность компрессии речевого сигнала по описанной выше схеме, но и открывает пути для создания новых методов энтропийного сжатия речевого сигнала (например, по схеме Шеннона-Фано), опирающихся на *априорное* знание о характере распределения лингвистических единиц в звучащем тексте. Используемые в настоящее время методы энтропийного сжатия обнаруживают низкую эффективность потому, что принимают в качестве исходных данных речевой сигнал вместе с шумом (фактически — не разделяя речевые и неречевые данные), вследствие чего распределение компрессируемых единиц оказывается неоптимальным и не поддаётся оптимизации стандартными способами.

ЗАКЛЮЧЕНИЕ

Произведённые в данной работе рассмотрения показали, что выделение лингвистической информации, содержащейся в речевом сигнале, способно повысить эффективность его сжатия на несколько порядков за счёт выведения большого количества информации в область статических данных кодека. В настоящее время этот подход не применяется, так как лишь сравнительно недавно необходимые для его реализации ресурсы (вычислительные мощности и память) стали достаточно дешёвыми для того, чтобы рассматривать описанную технологию как массовую. Между тем, необходимо заметить, что внедрение этой технологии в системах, реализующих постоянное хранение больших объёмов речевой информации, может дать значительную экономию ресурсов (прежде всего — долговременной памяти), поэтому разработка на её основе новых речевых кодеков представляется крайне перспективной.

[1] *Гуларян А. Б.* Электронный научно–образовательный журнал «Грани познания». Май 2010. № 1(6).
[2] *Абрамов Ю. В., Потапова Р. К., Хитина М. В.* Речевые технологии. № 3. 2010. С. 3.
[3] *Potapova R., Grigorieva M.* JECE. 2017.
[4] *Потапова Р.К., Потапов В. В.* Речевая коммуникация: от звука к высказыванию. М.: Издательский дом ЯСК, 2012.
[5] *Бобров Н.В.* Речевые технологии. 2017. № 1.
[6] *Потапова Р.К.* Речь: коммуникация, информация, кибернетика. 3-е изд. М.: Едиториал УРСС, 2003.
[7] *Zipf G. K.* Selected Studies of the Principle of Relative Frequency in Language. Cambridge. MA.: Harvard University Press, 1932.

УЗФФ № 5, 1750201 (2017)

Possibilities of utilizing linguistic information in development of advanced speech signal compression technologies

N. V. Bobrov

Moscow State Linguistic University. Moscow 119034, Russia
E-mail: arctangent@yandex.ru

Speech waveforms are known to be remarkably difficult to compress. Lossless entropy compression methods yield about 30% reduction of the initial waveform size. Delta-compression (and also somewhat similar to it methods based on the ideas suggested by A. Haar and I. Daubechies) allow to improve this result by another 30%. Lossy compression methods yielding much higher compression rates take advantage of the psychoacoustic peculiarities of sound perception by removing certain components of the speech signal that have the smallest effect over its perceived quality, e.g. portions of the spectrum that are «in the shadow» of prominent spectral peaks. It must be observed that all the abovementioned ideas are based on the existence of some a priori knowledge about such kind of data as the speech signal (for instance, that the ordinates of two adjacent points of the speech waveform normally differ by a small value, or that the spectrum of the speech signal usually contains several prominent peaks that determine the perceived quality of the sound). Following this logic, it can be suggested that the a priori knowledge that the speech signal is a container enclosing linguistic data can just as well be used to significantly improve the efficiency of both lossy and lossless compression technologies by including in the data source model the regularities appertaining to the linguistic component of the speech signal. This paper deals with the possible results of this suggestion being tested.

PACS: 43.72+q.
Keywords: speech signal, compression algorithms, codecs, linguistic information.
Received 26 June 2017.

Сведения об авторе

Бобров Николай Владимирович — науч. сотрудник; тел. (495) 637-56-97, e-mail: arctangent@yandex.ru.