Вопросы речевых технологий на XVI Международном конгрессе фонетических наук (2007 г.)

Е. В. Шаульский

В обзоре излагается содержание сообщений, представленных на XVI Международном конгрессе фонетических наук (Саарбрюккен, 2007 г.) в секции речевых технологий.

6–10 августа 2007 г. в Саарбрюккене (ФРГ) состоялся XVI Международный конгресс фонетических наук, в котором приняли участие 535 фонетистов из 39 стран. В настоящем обзоре рассматриваются доклады секции речевых технологий, представленные в сборнике трудов и материалов конгресса [1].

В работе *С. Крстулович, А. Хунекке и М. Шрёдера* (Саарбрюккен) [2] обсуждаются результаты использования скрытых марковских моделей (Hidden Markov Modelling, HMM) при синтезе экспрессивной немецкой речи. Авторы рассматривают преимущества системы синтеза, основанной на HMM, перед другими системами синтеза речи — формантно-ориентированными (formant-based) и конкатенативными (unit-selection based): в отличие от первых, HMM-системы обеспечивают высокое качество синтезируемого голоса, а в отличие от вторых, не так сильно зависят от лежащей в основе голосовой базы данных. В случае с экспрессивной речью важнейшим свойством системы синтеза является способность синтезировать просодические особенности речи. Для решения этой задачи авторы, во-первых, использовали базу данных нейтрального немецкого голоса — BITS German speech synthesis corpus, и, во-вторых, создали небольшую базу данных экспрессивных высказываний, имитирующих речь немецкого футбольного комментатора (Bundesliga database). Эти высказывания были подвергнуты параметризации с использованием скрытых марковских моделей, и их просодические характеристики были «наложены» на нейтральный голос из первой базы данных. Результаты эксперимента — синтезированные с помощью данной системы экспрессивные немецкие предложения — авторы оценивают как «в целом приемлемые», хотя и отмечают ряд недостатков их просодического оформления: несмотря на сохранение оригинального уровня тона, синтезированный экспрессивный голос звучит «сдавленно» и «скрипуче», темп синтезированной речи заметно ниже, чем в оригинале, а девиация

основной частоты (F0 deviation) у синтезированного голоса составляет всего 33 % от соответствующего параметра оригинала. В электронной версии Материалов конгресса к данной статье приложены аудиофайлы оригинальной и синтезированной немецкой речи, и читатель может самостоятельно оценить, насколько успешно используемая авторами система синтеза справляется со своей задачей.

Просодическому моделированию немецких слов посвящён и доклад *У. Хиршфельд, Р. Хоффмана и Ф. Ланге* (ФРГ) [3], которые занимаются созданием произносительного словаря немецкого языка (Aussprachwörterbuch), содержащего звуковой модуль. Система синтеза речи в таком словаре должна порождать произношение слов (заголовков словарных статей) на основании имеющейся фонетической транскрипции и приписанного каждому слову признака принадлежности к той или иной акцентной модели (включающей данные о соотношении длительностей гласных в слове, мелодическом контуре и т. п.). В докладе описан процесс совершенствования набора таких акцентных шаблонов для повышения качества синтезируемых слов.

М. Михкла (Таллинн) [4] указывает на значимость морфологических и синтаксических факторов в определении длительности сегментов при синтезе речи на эстонском языке. Автор произвёл моделирование длительности сегментов речи дикторов эстонского радио при помощи статистических методов (линейной регрессии и нейронных сетей), введя в исходные данные информацию о частеречной принадлежности слова, его синтаксической роли и морфологических признаках. Результатом этого стало уменьшение числа ошибок при предсказании длительности сегментов, что доказывает необходимость учёта не только фонетических, но и грамматических факторов при синтезе речи на языке с богатой морфологией (каким является эстонский).

К. Барткова и Д. Жуве (Ланьон, Франция) [5] рассматривают проблемы обнаружения иностранного акцента при автоматическом распознавании речи. Известно, что распознавание речи с иностранным акцентом является одной из наиболее сложных задач автоматического распознавания. Использование моделей, ориентированных на родной язык, не может достаточно хорошо справиться с речью иностранца; с другой стороны, модели, построенные на материале не только родного, но и иностранных языков, показывают худший результат в распознавании речи носителей языка. Авторы работы предлагают предварительно автоматически определить степень иностранного акцента, чтобы затем, с учётом полученных данных, применять ту или иную модель распознавания. Была создана база данных из французских слов, произнесённых носителями французского языка, а также носителями английского, немецкого и испанского языков. Для автоматического определения акцента производилось три цикла декодирования: в каждом случае использовалась контекстно-зависимая модель распознавания для французского языка, а также для одного из трёх других языков — немецкого, английского и испанского, после чего вычислялось соотношение сегментов, распознанных как французские, к общему их числу; в зависимости от величины этого коэффициента определялась степень иностранного акцента. Для распознавания «сильно акцентированной» речи в дальнейшем используется специальная модель, адаптированная для иностранного языка (foreign-adapted model). Если же степень акцента не очень высока, распознавание производится при помощи модели, ориентированной на родной язык (native model).

На более масштабном материале исследуют иностранные акценты во французском языке *Б. Виеру-Думулеску и её соавторы* (Орсе, Франция) [6]. Ими был создан корпус французских текстов, прочитанных носителями французского, арабского, английского, немецкого, итальянского, португальского и испанского языков (по 6 человек от каждого

языка). После этого было произведено измерение некоторых сегментных признаков, как то: частоты формант гласных, длительность согласных и степень их звонкости, а также наличие или отсутствие факультативного [ə] в финальной позиции. Затем для каждого слова были определены произносительные варианты с учётом фонетических особенностей того или иного акцента, например, оглушения звонких, фрикативизации смычных, различных реализаций /r/, назальных гласных и т. п., и создан своего рода словарь произносительных вариантов. В результате выявились произносительные «предпочтения» носителей того или иного языка, говорящих по-французски (арабы не различают /e/ и /i/, немцы и англичане произносят глухие согласные с придыханием, испанцы не делают различия между /b/ и /v/, и т. п.) и определена их частотность.

Ф. Була де Мареуиль, М. Адда-Деккер и С. Вёрлинг (Орсе, Франция) [7] исследовали реализацию ртовых и носовых гласных в северной и южной разновидностях французского языка. Для этого они создали корпус данных из записей речи 12-ти географических пунктов северной и южной Франции. Обследование этого корпуса (с использованием методов автоматической обработки речи) позволило получить количественные данные о реализации гласных фонем, подтвердившие давние наблюдения французских диалектологов и социолингвистов: более переднее произношение /ё/ (вплоть до [œ]) характерно для северной Франции, тогда как на юге более частотно расщепление носовых гласных на (носовой или ртовый) гласный + носовой согласный.

Доклад *И. Лапри и А. Бонно* (Франция) [8] посвящён построению стимулов перцепции при помощи «синтеза с копированием» (copy synthesis). Для экспериментов по восприятию требуются речевые стимулы с изменяемым акустическим содержанием. Одну из возможностей для создания таких стимулов предоставляет система формантного синтеза, позволяющая вручную редактировать параметры синтезируемого звука. Авторами доклада была предложена система «синтеза с копированием», развивающая возможности формантного синтезатора. Синтез с копированием включает два этапа. Вначале производится вычисление основной частоты и определение параметров источника, в частности, соотношение голоса и шума. Второй шаг — задание формантных амплитуд. В данной работе намечаются пути развития системы синтеза с копированием в двух направлениях: автоматическое отслеживание динамики частоты формант (automatic formant tracking), при котором учитывается взаимозависимость формантных кривых, а также задание формантных уровней с использованием алгоритма тонального маркирования.

Проект системы синтеза речи по артикуляционным данным «Ouisper» представлен в работе *Т. Уэбера, Ж. Шолле, Б. Денби, М. Стоун и Л. Зуари* (Париж — Балтимор) [9]. Синтез речи в системе «Ouisper» должен осуществляться на основании артикуляционных данных, полученных из ультразвуковых изображений речевого тракта и видеозаписей движения губ говорящего. Применение HMM-моделирования к корпусу аудиовизуальных данных (также использовался алгоритм Unit-Selection) позволяет соотнести акустические данные с артикуляционными и построить систему синтеза речи, которая может быть использована в качестве альтернативы трахео-пищеводной речи больных раком гортани, в ситуациях, требующих сохранения тишины, а также для голосового общения в обстановке повышенного шума. Обсуждаются проблемы предварительной обработки ультразвуковых изображений, извлечения релевант-

ной информации из положения языка и губ, автоматической сегментации акустического сигнала. Авторы утверждают, что на данном этапе система продуцирует фонетическую транскрипцию только на основании видеосигнала (т. е. без обращения к аудиоданным) с точностью 50 %. В дальнейшем для решения задачи синтеза предстоит разработать систему «виртуальной просодии», для чего предполагается использование автоматизированной модели извлечения «просодических шаблонов» из данных корпуса.

Группа исследователей из Германии, Италии, Франции и Израиля [10] представила доклад, посвящённый проблемам соотношения тона и длительности в эмоциональной и аффектированной речи. Ставится под сомнение традиционная точка зрения, что тон играет более важную роль в маркировании эмоциональных состояний, чем другие просодические признаки. Авторы доклада провели следующее исследование. Была создана речевая база данных высказываний детей при общении с собакой-роботом AIBO. Полученные высказывания затем классифицировались как нейтральные либо содержащие какую-либо эмоцию (злость, нежность, эмпатия). После этого при помощи системы ESPS было осуществлено автоматическое извлечение данных о движении частоты основного тона в этих высказываниях, после чего полученные данные были скорректированы вручную одним из авторов на основе принципа «сглаживания и адаптации к человеческому восприятию», для того чтобы исключить влияние модуляций фонации (скрипучий голос, ларингализация и т. п.) на тональный контур. Таким образом, были получены два корпуса данных: первый — из автоматически определённых значений тонального движения (*aut*), второй — из тех же значений, подвергшихся «ручной» коррекции (*corr*). Для каждого из этих корпусов было вычислено соотношение параметров тона (F0) и длительности (DUR) в определении принадлежности высказывания к той или иной эмоции. Оказалось, что параметр F0 для *aut* более существен, чем для *corr*, тогда как DUR играет более важную роль для *corr*, чем для *aut*. Вместе с тем это различие не является «отчётливо выраженным». Авторы не дают ответа на вопрос, свидетельствуют ли полученные ими результаты о меньшей значимости тонального фактора в эмоциональной речи или же лишь об ошибках в автоматическом определении высоты тона, однако отмечают, что их выводы подчёркивают важность параметра длительности в общем комплексе просодических признаков.

Э. Лазарчик (Саарбрюккен) [11] посвятила сообщение изучению положений гортани для задач синтеза речи в артикуляторной модели. Ею было исследовано влияние положения гортани на качество гласных: во-первых, при подъёме гортани повышаются формантные частоты гласного; во-вторых, от положения гортани зависит и качество голоса: более высокое положение гортани соответствует напряжённому голосу, более низкое — расслабленному. Далее изолированные гласные в естественном произношении, но произнесённые с разным положением гортани — нейтральным, повышенным и пониженным, — сравнивались с гласными, синтезированными при помощи трёхмерной артикуляционной модели вокального тракта, где положение гортани было соответствующим образом симулировано. Проведённые измерения частоты первых трёх формант естественных и синтезированных гласных показали, что изменение положения гортани влияет на частоты формант и в том, и в другом случае. Что касается качества голоса, то манипулирования высотой гортани в артикуляторной модели оказалось недостаточно, чтобы достичь характеристик, присущих естественной речи. Помимо этого, потребовалось изменение параметров возбуждения, которые в сочетании с положением гортани позволяют достичь искомого результата.

Ф. Алиас и М. П. Тривиньо (Барселона) [12] предлагают таблицу для оценки разборчивости речи на каталанском языке. Описана процедура построения сбалансированной таблицы с учётом частотности согласных фонем в каталанском языке. В результате сформирована

таблица из 40 четырёхсловных списков, из которых 20 построены на изменении последней согласной фонемы, а другие 20 — на изменении первой.

В докладе *Дж. Уэллса* (Лондон) [13] обсуждаются вопросы использования фонетических символов в различных компьютерных приложениях (текстовых редакторах, программах электронной почты, веб-страницах и пр.). За последние годы широко распространился единый формат кодирования символов — Юникод, что позволяет и для фонетических символов применять единые методы кодирования, отказываясь от разнообразных шрифтов, не соответствующих международному стандарту. Одной из задач остаётся облегчение клавиатурного ввода специальных символов, для чего имеется ряд способов: Alt+номер, Таблица символов, Alt+X, специальные раскладки клавиатуры.

Литература

1. ICPhS 2007 — Proceedings of 16th International Congress of Phonetic Sciences (6–10 August 2007, Saarbrücken, Germany) // Edited of Jürgen Trouvain and William J Barry. Saarbrücken, 2007. Электронная версия: http://www.icphs2007.de.

2. Krstulović S., Hunecke A., Schröder M. Investigating HMMs as a parametric model for expressive speech synthesis in German // ICPhS 2007. P. 2181–2184.

3. Hirschfeld U., Hoffmann R., Lange F. Prosodic modelling of synthesised German words // ICPhS 2007. P. 2205–2208.

4. Mihkla M. Morphological and syntactic factors in predicting segmental durations for Estonian text-to-speech synthesis // ICPhS 2007. P. 2209–2212.

5. Bartkova K., Jouvet D. Automatic detection of foreign accent for automatic speech recognition // ICPhS 2007. P. 2185–2188.

6. Vieru-Dumulescu B., Boula de Mareüil Ph., Adda-Decker M. Characterizing non-native French accents using automatic alignment // ICPhS 2007. P. 2217–2220.

7. Boula de Mareüil Ph., Adda-Decker M., Woehrling C. Analysis of oral and nasal vowel realisation in Northern and Southern French varieties // ICPhS 2007. P. 2221–2224.

8. Laprie Y., Bonneau A. Construction of perception stimuli with copy synthesis // ICPhS 2007. P. 2189–2192.

9. Hueber T., Chollet G., Denby B., Stone M., Zouari L. Ouisper: Corpus based synthesis driven by articulatory data // ICPhS 2007. P. 2193–2196.

10. Batliner A., Steidl S., Schuller B., Seppi D., Vogt T., Devillers L., Vidrascu L., Amir N., Kessous L., Aharonson V. The impact of F0 extraction errors on the classification of prominence and emotion // ICPhS 2007. P. 2201–2204.

11. Lasarcyk E. Investigating larynx height with an articulatory synthesizer // ICPhS 2007. P. 2213–2216.

12. Alías F., Triviño M. P. A phonetically balanced modified rhyme test for evaluating Catalan speech intelligibility // ICPhS 2007. P. 2197–2200.

13. Wells J. An update on phonetic symbols in Unicode // ICPhS 2007. P. 2225–2228.

Е.В. Шаульский
Аспирант филологического факультета МГУ им. М. В. Ломоносова.