УДК 621.865.8: 534.78

ПРОБЛЕМЫ РАСПОЗНАВАНИЯ СЛИТНОЙ РЕЧИ

Леонович А.А.

Описываются характеристики систем распознавания речи и их классификация. Обсуждаются методы сегментации и проблемы распознавания речи. Приводятся сравнительные характеристики методов сегментации.

Введение

Голосовой интерфейс это тема, которая на протяжении последних пятидесяти лет привлекает внимание ученых и инженеров всего мира. Голосовой интерфейс на языке пользователя – это наилучшее решение, поскольку речь – это наиболее натуральная, удобная, эффективная и экономичная форма человеческого взаимодействия.

Речевой ввод объединяет множество различных технологий и приложений. В некоторых случаях важно не понимание лингвистического содержания, а идентификация говорящего или языка, на котором происходит общение. Идентификация говорящего в свою очередь может включать в себя получение некоторых специфических параметров, определяющих данного пользователя: пола, возраста или местности проживания по характерному диалекту.

Однако в основе технологии речевого ввода в первую очередь лежит процесс распознавания речи. В соответствии с общепринятым определением, распознавание речи – это процесс преобразования акустического сигнала, полученного с микрофона или телефона, в набор слов. Для таких приложений как командные системы, системы ввода данных или обработки документации, распознанные слова могут сами по себе быть конечным результатом. Однако в некоторых случаях, полученные после распознавания данные могут подвергаться дальнейшей лингвистической обработке для получения форматированного текста либо для достижения понимания речи машиной.

Одним из наиболее сложных аспектов в разработке систем машинного распознавания речи является широкая междисциплинарность задачи. При работе над такими системами затрагиваются вопросы теории обработки сигналов, математического анализа, лингвистики, теории коммуникаций, физиологии и в некоторых случаях психологии. И для построения успешной системы распознавания необходимо рассмотреть такой круг дисциплин, который один человек охватить не в состоянии. Следовательно, для разработчика становится особенно важно понимание основ речевого распознавания без необходимости быть экспертом в каждой из сторон проблемы.

Распознавание речи - это не простая задача, особенно усложненная тем, что существует множество источников вариативности, связанных с речевым сигналом:

1. Акустическая реализация фонем, как наименьших звуковых единиц, составляющих слово, в большой степени зависит от контекста, в котором фонема появляется. В потоке речи звуки видоизменяются под влиянием соседних фонем, причем иногда звуки накладываются друг на друга или вообще выпадают: *говорить* — [гъвар'ит'].

2. Акустические изменения могут быть вызваны влиянием окружающей среды, а так же характеристиками и позицией приемника речевого сигнала.

3. Физическое и эмоциональное состояние диктора, темп или качество произношения также могут вносить свой вклад в изменчивость речи.

4. Различия в социолингвистическом окружении, диалекте и объеме речевого тракта способствуют возникновению междикторской вариативности.

Классификация систем распознавания речи

Системы распознавания речи могут характеризоваться множеством параметров:

– по типу речи различают системы распознавания изолированных слов и слитной речи;

– по стилю речи: речь может быть спонтанной, либо зачитанной;

– по типу диктора: дикторозависимые и дикторонезависимые;

– по размеру словаря: с маленьким (< 100 слов) или большим словарем (> 10 000 слов).

Системы распознавания изолированных слов требуют от пользователя внесения кратких пауз в речи между словами. Такие системы давно себя зарекомендовали на коммерческом рынке подобных продуктов и показывают достаточно высокие результаты, в частности, в составе командных систем [1]. Однако, данный способ речевого ввода уже является не натуральным для человека, так как требует искусственного изменения потока речи. Это приводит к замедлению процесса ввода информации и быстрой усталости диктора. Соответственно, системы распознавания слитной речи являются более предпочтительными, хотя создание подобных систем - задача во много раз более сложная.

Особые условия на процесс распознавания слитной речи накладывает ее стилистика – это может быть зачитанный текст научной статьи, художественное произведение или часть спонтанного диалога. В данном случае обработка спонтанной речи будет требовать значительно больших усилий и ресурсов, так как такая речь носит произвольный характер, часто не подчиняется правилам языка и полна эмоциональной окраски и междометий [2].

Некоторым системам перед началом работы с новым пользователем необходимо пройти предварительный этап обучения, т.е. каждый диктор должен предоставить образцы своей речи для дальнейшей работы. Это делает подобные системы дикторозависимыми, что в некоторой степени ограничивает и усложняет их использование. Но в то же время, это позволяет повысить качество распознавания, а в ряде случаев данное ограничение может выступать в качестве меры безопасности, осуществляя контроль доступа.

При проектировании систем распознавания речи важно учитывать область применения: либо это будет система диктовки текста, работающая со словарем, содержащим несколько тысяч записей, либо командная система, обрабатывающая несколько десятков слов. Распознавание для систем с большим словарем, многие из слов которого могут звучать достаточно схоже, требует применения особых комплексных алгоритмов поиска и обработки информации.

Сейчас можно сказать, что наиболее привлекательным, с точки зрения пользователя, становится создание дикторонезависимой системы распознавания слитной речи, работающей с большим словарем, при этом рассчитанной на обработку спонтанной речи. Однако, для разработчика данная проблема является наиболее трудной в силу вышеизложенного. Но не смотря на это, создание подобной системы в настоящее время является первостепенной задачей в теории речевого ввода.

Речевая единица сегментации

Деление непрерывной речи на элементарные единицы это одна из наиболее сложных задач в процессе распознавания слитной речи. Обычно, данная проблема подразделяется на две независимые подзадачи:

- преобразование речевого сигнала в строку дискретных минимальных единиц речи (фонем) с последующей классификацией;
- разделение полученной строки на значимые сегменты (слова, или в более общем смысле, лексические единицы).

Для большинства языков, первая задача уже является достаточно трудной, в частности, из-за коартикуляции, когда происходит взаимовлияние соседних звуков при произнесении. Звуки накладываются, создавая переходные участки как внутри слов, так и на стыках смежных, что становится особенно проблематичным для второй подзадачи. Примером могут служить фразы: *Про силу вы ли говорили? – Просил, увы ли говорили...* Поскольку в нормальной разговорной речи не существует заметных пауз между словами, становится достаточно не просто правильно выделить лексемы.

Сейчас большинство систем распознавания слитной речи с большим словарем для моделирования элементарной акустической единицы используют дифон или его контекстно-зависимый вариант – трифон [3-5]. Такой выбор обусловлен тем, что дифон – это звуковая единица, имеющая протяженность от середины одного звука до середины последующего. Считается, что речевой сигнал содержит стационарные участки звуков, независящие от влияния соседних, т.е. не подверженные коартикуляционному эффекту. В середине такого участка и проводится граница дифона. Следовательно дифон сохраняет информацию, хранящуюся в переходном участке между фонемами, которая, как было доказано, является полезной [6].

Однако, универсальность использования дифонов для представления речи также можно поставить под вопрос, поскольку эти единицы часто не отражают все сложности речевого сигнала. Во-первых, коартикуляционные эффекты обычно широко растянуты по времени и соответствующие временные зависимости не могут быть переданы дифонами, которые носят характер кратковременного сегмента. Во-вторых, само использование дифонов основано на упрощении, что слово состоит из набора фонем. В рамках такого представления различия в произношении могут быть выражены только заменой, вставкой или удалением фонем. Более того, такое описание ограничивает возможности эффективного использования фонетических зависимостей более высокого порядка, например, относящихся к слоговой структуре слова [7].

Одним из возможных решений данной проблемы может быть выбор акустической единицы, которая содержит спектральную и временную информацию в себе. Наиболее очевидными кандидатами для этих целей кажутся слово и слог. В частности, правильность использования слога подтверждается исследованиями в области фонетики и восприятия речи человеком [8,9]. У слога существует так называемая *произносительная неделимость*, доказанная экспериментально – как бы ни была замедленна речь, она никогда не распадается на промежутки меньшие чем слог.

Тем не менее, выбор более длинной единицы сегментации речевого сигнала для распознавания речи не является чем-то новым: модели, использующие слово как минимальный сегмент, широко используются в приложениях с ограниченным словарем, таких как распознаватели цифр или набора команд. Использование слоговых моделей также предлагалось ранее [10,11]. Но подобный подход приводит к возникновению новой проблемы при работе с системами с большим словарем – это недостаток данных для обучения системы, создания начальных эталонов. С повышением длины элементарной единицы речи возрастает количество вариантов данного элемента. И в случае выбора слова в качестве такой меры, число вариаций увеличивается до недопустимых пределов. Применительно к русскому языку, существуют разработки, в которых применен метод деления речи на морфемы, что значительно уменьшает размер словаря. Однако, вследствие того, что морфема не является акустической единицей, данный способ требует дополнительных усилий в плане построения морфем из сегментов, полученных после обработки речевого сигнала, что в свою очередь снижает эффективность алгоритма [12]. Существуют разработки, которые пытаются преодолеть данную проблему, путем совместного иерархического использования нескольких видов речевых сегментов: слов, слогов и фонем [13,14]. И все же, даже в случае применения таких иерархических систем, задача однозначной сегментации речевого сигнала все еще окончательно не решена.

Методы сегментации речевого сигнала

Прежде чем непосредственно приступить к распознаванию речи, в первую очередь необходимо создать базу эталонов речевых единиц. И здесь снова встает задача сегментации речевого сигнала. Самым первым и наиболее простым способом получения элементарных единиц явля-

лась ручная обработка записанных фраз. Эксперты выполняли сегментацию основываясь на спектрограммах, кривых энергии, интонациях и других приемах, используемых в речевом анализе. Этот способ обладает некоторыми преимуществами перед автоматическими методами – опытные лингвисты могут с большой точностью, анализируя многие факторы, определить границы сегментов. Но в то же время, такая процедура очень трудоемка и требовательна к ресурсам, что делает данный метод применимым только в ограниченных случаях.

Одним из основных направлений в области распознавания слитной речи является применение аппарата скрытого Марковского моделирования (СММ), где речевой сигнал представляется набором состояний с некоторыми вероятностями перехода между ними. Отсюда возник метод автоматической сегментации речи.

Рассмотрим схему системы распознавания слитной речи на основе СММ, работающей со звуковыми единицами, длиной меньше слова (рис. 1). Этап построения модели слова по сути является некоторым вариантом сегментации – здесь происходит подбор эталонов для входного вектора признаков. Соответственно, выполнив некоторые модификации для уменьшения вычислительной сложности полного процесса распознавания, данный процесс можно применить для автоматического поиска границ элементарных единиц речи [15].

Однако, классический процедура, основанная на СММ, требует полного транскрибирования входного потока речи, другими словами необходимо пройти полный процесс речевого распознавания.

Другой класс алгоритмов рассматривает речь только с позиций обработки акустических сигналов. В простейшем случае, происходит разделение речевого тракта на сегменты, где в качестве границ выступают паузы [16]. Различные характеристики служат для определения границ данной речевой единицы в речевом тракте: энергетические параметры [17-19], резонансные частоты и частота основного тона [20], а также разнообразная просодическая информация [21].

Существуют и иные альтернативные подходы к сегментации речи, не обязательно ограниченные техникой речевой обработки, но и применяющие обобщенные статистические алгоритмы. Среди них: нейросетевой метод [22, 23], статистическое моделирование [24] и динамическое программирование [25,26] (таблицу 1).

Сегментация речевых баз данных – одна из наиболее актуальных задач. В этом случае соответствующая транскрипция обычно известна заранее, и быстрые алгоритмы обработки не являются необходимыми. С другой стороны, если сегментация является частью процесса распознавания речи, то при отсутствии какой либо информации о лексическом содержании сигнала, необходима обработка в реальном времени. Следовательно, характеристики подходящего алгоритма должны определяться исходя из задач итогового приложения.

Заключение

Следует сказать, что при всем многообразии существующих подходов и методов добиться максимально высоких показателей так и не удалось. Хотя исследования и эксперименты в области речевого распознавания проводятся уже не один десяток лет, ясно, что все еще остается множество нерешенных проблем. Так и не решена задача построения распознавателя, обладающего характеристиками на уровне человеческого восприятия речи.

Однако уже сейчас можно обозначить первостепенные проблемы в теории распознавания речи и, в частности, для ее непрерывного варианта:

1. Выбор наиболее подходящей речевой единицы, для сегментации и представления речевого потока. Это особенно актуально для систем, работающих с большим словарем.

2. Построение автоматического метода поиска границ сегментов речевых единиц в потоке слитной речи. Существующие техники все еще в большой степени уступают способу ручной сегментации.

Таблица 1

Характеристики методов сегментации речи (по отношению к ручной разметке, которая считалась эталонной)

Метод сегментации	Метод СММ [15]	Метод спектр. характеристик [18]	Нейросетевой метод [23]	Метод стат. моделирования [24]	Байесовский метод предсказаний [27]	Совместный метод [28]
% правильных границ (допуск ±15 мсек)	85.5%	85%	83%	84.5%	-	-
% правильных границ (допуск ±20 мсек)	-	-	94.19%	89.51%	76.21%	90.22%

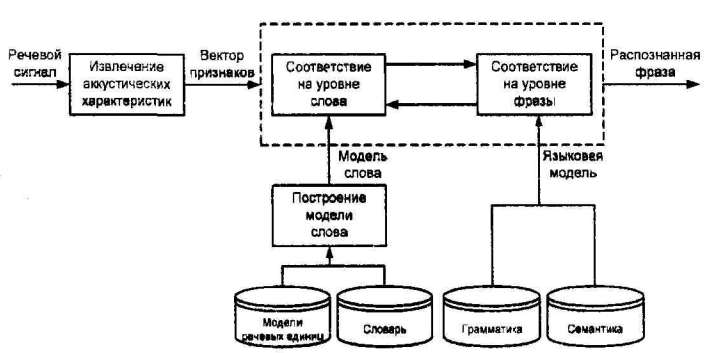

Рис. 1 Схема системы распознавания слитной речи

В рамках работ в области сегментации речи проводятся эксперименты в плане применения аппарата вейлет-преобразования для поиска границ речевых сегментов [30]. Последние результаты показали, что наиболее удобной единицей сегментации является слог, так как. информация о фонемах распределена на всем протяжении данной единицы. Полученные данные хорошо согласуются с выводами исследователей кафедры фонетики Санкт-Петербургского Университета [6,7]. Это, в свою очередь, позволило сделать вывод о целесообразности проведения дальнейших экспериментов именно в направлении слоговой сегментации непрерывного потока речи, что сейчас и осуществляется.

Литературы

1. Леонович А.А. Современные технологии распознавания речи // Международный семинар Диалог'2005 по компьютерной лингвистике и ее приложениям. Электронная публикация на сайте конференции. www.dialog-21.ru
2. Weintraub M., Taussig K., Hunicke-Smith K. & Snodgrass A. Effect of Speaking Style on LVCSR Performance. Proc. Int. Conf. on Spoken Language Proc., supplement, 1996.
3. Watanabe, S., Sako, A., Nakamura, A. Automatic determination of acoustic model topology using variational Bayesian estimation and clustering for large vocabulary continuous speech recognition. Audio, Speech and Language Processing, IEEE Transactions on. Volume 14, Issue 3, May 2006, pp. 855-872.
4. Sakti, S., Markov, K., Nakamura, S. Incorporation of Pentaphone-Context Dependency Based on Hybrid Hmm/Bn Acoustic Modeling Framework. Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. Volume 1, 14-19 May 2006, pp. 1177-1180.
5. Jurafsky D., Ward W., Jianping Z., Herold K., Xiuyang Y., and Sen Z. What kind of pronunciation variation is hard for triphones to model? Proc. ICASSP-2001, Salt Lake City, Utah, USA, May 8-11. 2001, vol. I, pp. 577-580.
6. Бондарко Л.В., Кузнецов В. И., Скрелин П.А., Шалонова К. Б. Звуковая система русского языка в свете задач компилятивного синтеза // Бюллетень фонетического фонда русского языка. № 6, май 1997.
7. Белявский В. М., Светозарова Н. Д. Слоговая фонетика и три фонетики Л.В. Щербы. Статья, расположенная по адресу http://www.auditech.ru/doc/cherba.htm
8. Schiller N.O., Meyer A.S., and Leveit W.J.M. The syllabic structure of spoken words: Evidence from the syllabification of intervocalic consonants. Language & Speech, vol. 40, 1997, pp. 103-140.
9. Зиндер Л.Р. Общая фонетика. М., 1979.
10. Greenberg S. Speaking in shorthand – a syllable-centric perspective for understanding pronunciation variation. Speech Communication, vol. 29, pp. 159-176, 1999.
11. Jones R.J., Downey S., and Mason J.S. Continuous speech recognition using syllables," in Proc. Eurospeech-97, Rhodes, Greece, Sept 22-25. 1997, vol. 3, pp. 1171-1174.
12. Ронжин А.Л., Карпов А.А., Ли И.В.Система автоматического распознавания русской речи SIRIUS.// Санкт-Петербургский институт информатики и автоматизации РАН.
13. Hämäläinen, A., de Veth, J., and Boves, L. Longer-Length Acoustic Units for Continuous Speech Recognition, in Proc. EUSIPCO-2005, Antalya, Turkey, Sep 4-8, 2005.
14. Messina, R. and Jouvet D. Context dependent "long units" for speech recognition, in Proc. ICSLP-2004, Jeju Island, Korea , Oct 4-8, 2004, pp. 645-648.
15. Brugnara F., Falavigna D., and Omologo M. Automatic segmentation and labeling of speech based on hidden Markov models. Speech Communication, vol. 12, no. 4, pp. 357-370, 1993.
16. Pfeiffer S. Pause Concepts for audio Segmentation at Different Semantic Levels. ACM Multimedia, 2001, pp. 187-193.
17. Milone D.H., Merelo J.J., Rufiner H.L. Evolutionary algorithm for speech segmentation. Evolutionary Computation, 2002. CEC '02, vol. 2, 12-17 May 2002, pp. 1115-1120.
18. Сорокин В.Н., Цыплихин А.И. Сегментация и распознавание гласных. Информационные процессы, том 4, № 2, 2004, стр. 202-220.
19. Ермоленко Т.В., Шевчук В.В. Алгоритмы сегментации с применением быстрого вейвлет-преобразования. // Международный семинар Диалог'2003 по компьютерной лингвистике и ее приложениям. Электронная публикация на сайте конференции. www.dialog-21.ru
20. Wendt C., Petropulu A.P. Pitch determination and speech segmentation using the discrete wavelet transform. Circuits and Systems, 1996. ISCAS '96, 'Connecting the World', vol. 2, 12-15 May 1996, pp. 45 – 48.
21. Dong Wang, Lie Lu, Hong-Jiang Zhang. Speech segmentation without speech recognition. Multimedia and Expo, 2003. ICME '03. Proceedings, vol. 1, 6-9 July 2003, pp. 405-408.
22. Vorstermans A., Martens J.-P. and Van Coile B. Automatic segmentation and labeling of multi-lingual speech data. Speech Communication, vol. 19, pp. 271-293, 1996.
23. Toledano D.T. Neural network boundary refining for automatic speech segmentation. Acoustics, Speech, and Signal Processing, 2000. ICASSP '00, vol. 6, 5-9 June 2000, pp. 3438-3441.
24. Pauws S., Kamp Y. and Willens L. A hierarchical method of automatic segmentation for synthesis applications. Speech Communications, vol. 20, pp. 207-220, 1996.
25. Bajwa R.S., Owens R.M., Kelliher T.P. Simultaneous speech segmentation and phoneme recognition using dynamic programming. Acoustics, Speech, and Signal Processing, 1996. ICASSP-96. Conference Proceedings, vol. 6, 7-10 May 1996, pp. 3213-3216.
26. Sharma, M., Mammone R. "Blind" speech segmentation: automatic segmentation of speech without linguistic knowledge. Spoken Language, 1996. ICSLP 96. Proceedings., Fourth International Conference on, vol. 2, 3-6 Oct. 1996, pp. 1237-1240.
27. Ming Liu, Huang T.S. A Bayesian Predictive Method for Automatic Speech Segmentation. Pattern Recognition, 2006. ICPR 2006. 18th International Conference on, vol. 4, 20-24 Aug. 2006, pp. 290-293.
28. Runqiang Yan, Yiqing Zu, Yisheng Zhu. Automatic Speech Segmentation Combining an HMM-Based Approach and Recurrence Trend Analysis. Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings, vol. 1, 14-19 May 2006, pp. 797-800.
29. Rabiner L. and Juang B.-H. Fundamentals of Speech Recognition, Prentice Hall, Englewood Cliffs, New Jersey, 1993.
30. Леонович А.А. Выбор вейвлет базиса для алгоритма автоматической сегментации речи. Издательство УРСС. Коллектив авторов. Первая Международная конференция «Системный анализ и информационные технологии» САИТ-2005 (г., Переславль-Залесский, Россия): Труды конференции. 2005.