УДК 004.896

МЕТОД АВТОМАТИЧЕСКОГО РАСПОЗНАВАНИЯ ГОЛОСОВЫХ КОМАНД И НЕРЕЧЕВЫХ АКУСТИЧЕСКИХ СОБЫТИЙ

А. Л. Ронжин,
доктор техн. наук, доцент
С. В. Глазков,
аспирант
Санкт-Петербургский институт информатики и автоматизации РАН

Рассматривается метод анализа акустических данных, позволяющий классифицировать по голосовым командам пользователя и неречевым звукам текущую ситуацию в помещении и принять меры в случае возникновения чрезвычайных обстоятельств. Разработанный метод распознавания акустических элементов основан на применении математического аппарата скрытых марковских моделей.

Ключевые слова *— автоматическое распознавание речи, цифровая обработка сигналов, ассистивные системы, интеллектуальное пространство.*

Введение

Разработка проактивных сервисов и многомодальных интерфейсов, обеспечивающих бесконтактный ненавязчивый способ взаимодействия с автоматизированными системами жизнеобеспечения и безопасности, сейчас особенно актуальна в связи с наблюдаемой тенденцией увеличения процента граждан пожилого возраста и людей с ограниченными возможностями [1]. За рубежом активно разрабатываются так называемые ассистивные системы, ориентированные на улучшение комфорта, безопасности и повышение здоровья и независимости людей [2, 3]. Системы, связанные с обеспечением комфорта, направлены на автоматизацию и увеличение дружественности интерфейсов стандартного бытового оборудования. Скрининговые системы мониторинга здоровья человека определяют текущее состояние человека на основе данных физиологических сенсоров: частоты пульса, веса и других параметров. Системы безопасности следят за возникновением чрезвычайных ситуаций, например пожара, протечки, взлома, а также падения человека, стонов, плача.

Технологии автоматической обработки звуковой информации играют важную роль в обеспечении и безопасности жизнедеятельности, автоматизации управления бытовым оборудованием и мониторинге текущего состояния человека [4]. Система определения падения звука человека, отличающаяся использованием метода трехмерной локализации источника звука, описана в работе [5]. Диалоговая система, включающая модули автоматического распознавания и синтеза речи, успешно применяется [6] для определения физических недомоганий или снижения когнитивных способностей на основе анализа речевых ответов пользователя. Анализ видеоданных совместно с неречевыми звуковыми данными используется в проекте Sweet Home [7] для локализации пользователя в помещении.

Существующие методы анализа акустических элементов

Хотя речь является наиболее информативным акустическим событием, другие типы звуков, произведенные человеком или какими-либо объектами, также несут полезную информацию для ассистивных интеллектуальных систем. Распознавание неречевых акустических событий, например аплодисментов, смеха, кашля, перемещения кресла, скрипа двери и т. д., может помочь в определении и описании текущей активности человека. Кроме того, за счет определения неречевых звуков и их исключения из аудиосигнала можно увеличить точность системы автоматического распознавания речи.

Первые системы идентификации акустических событий обладали небольшим словарем для обработки ограниченного набора звуков. В боль-

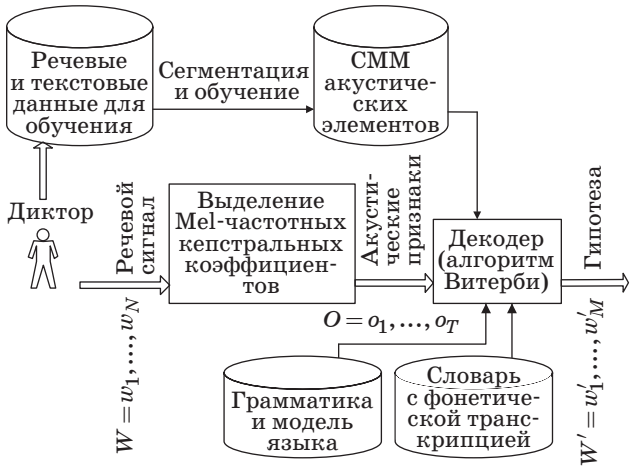
шинстве систем при параметрическом представлении звукового сигнала используются классические методы цифровой обработки речи: метод кепстральных коэффициентов на шкале мел (MFCC), метод коэффициентов перцептивного линейного предсказания (PLP), а при распознавании — скрытые марковские модели (СММ) или модели гауссовых смесей. В работе [8] авторы сравнивают между собой различные признаки и классификаторы, относящиеся к таким повседневным контекстам, как ресторан, машина, библиотека и офис. Система использует метод MFCC и их первую производную по времени, а также СММ. В работе [9] представлена система определения акустических событий в записях совещаний. Модели событий созданы с использованием сети СММ, их размер и топология выбраны на основе обучения распознавания изолированных событий. Авторами также учитывается эффект влияния внешнего шума на точность идентификации акустических событий.

Эксперименты [10] показали, что в среднем люди способны распознавать повседневный аудиоконтекст в 70 % случаев, и большинство ошибок распознавания возникают между контекстами одних типов. Также было показано, что распознанные отдельные акустические события из звукового сигнала для человеческого восприятия аудиоконтекста являются отдельной репликой. Однако большинство предложенных систем распознавания контекста моделируют глобальные акустические характеристики аудиоконтекста, а не акустические события [11]. В работе [10] авторы представляют систему распознавания контекста, основанную на определении индивидуальных акустических событий, а также метод, основанный на предположении, что различные контекстные пространства, такие как улица или ресторан, характеризуются возникновением определенных акустических событий. Контекст смоделирован с использованием гистограмм событий, собранных из аннотированных записей. Работа системы разделена на два этапа: определение звукового события и распознавание контекста. На первом осуществляется выделение событий, присутствующих в тестируемом контексте. Гистограмма событий, построенная по результату работы системы, приводится в соответствии с моделями контекста. Оценка системы осуществляется с помощью десяти контекстов, которые включают в себя одинаковые события. Средняя точность классификации акустических событий составила 89 % [10]. Объединение системы распознавания контекста с обычной системой распознавания аудиосигнала позволило повысить точность ее работы до 92 %.

Метод распознавания акустических событий в ассистивном интеллектуальном пространстве

При разработке ассистивного многомодального интеллектуального пространства использовалось многоканальное оборудование записи и обработки видео- и аудиосигналов [12, 13]. Схема работы метода распознавания акустических элементов, предусматривающая два режима [4]: обучение моделей и декодирование речи, — показана на рисунке.

При разработке системы распознавания акустических событий в ассистивном интеллектуальном пространстве был составлен лексический словарь, включающий четыре голосовые команды: множество V={“yes”, “no”, “help”, “trouble”} и тринадцать типов акустических событий: множество E={“throatcleaning”, “cry”, “cough”, “applause”, “fall”, “step”, “keydroping”, “keyjingle”, “chairmoving”, “door”, “paperwork”, “phonering”, “water”}, по которым можно судить о текущей обстановке в помещении. Неречевые акустические элементы также следует разделить на артефакты речи; звуки, издаваемые человеком в процессе жизнедеятельности; звуки, возникающие в процессе взаимодействия человека с предметами, а также искусственные или естественные звуки, но не связанные с человеком, т. е. звуки других объектов. В табл. 1 представлена классификация анализируемых акустических элементов по перечисленным группам. Также вводится множество тревожных акустических событий, состоящее из следующих речевых и неречевых элементов: E={“moan”, “cough”, “fall”, “keydroping”, “help”, “trouble”}, — появление которых чаще всего случается в чрезвычайной ситуации. Данное множество может быть расширено в зависимости от условий, где применяется система распознавания речи, и особенностей жизнедеятельности пациентов.



■ Схема работы метода распознавания акустических элементов

■ ***Таблица 1.*** Классификация анализируемых акустических элементов

Название элемента	Обозначение	Принадлежность к множеству					
		голосовых команд	неречевых элементов				тревожных акустических событий
			Артефакты	Звуки, издаваемые человеком	Звуки при взаимодействии человека с предметами	Звуки других объектов	
Прочищение горла	th		+				
Стон	mo		+				+
Кашель	co		+				
Аплодисменты	ap			+			
Падение	fa			+			+
Шаги	st			+			
Звон ключей	kd				+		
Падение ключей	kj				+		+
Перемещение кресла	cm				+		
Стук двери	ds				+		
Шелест бумаги	pw				+		
Звонок телефона	pr					+	
Течение воды	wa					+	
Да	ys	+					
Нет	no	+					
Помогите	hl	+					+
Проблема	tr	+					+

Для обучения моделей акустических элементов был собран корпус из 2600 аудиозаписей общей длительностью 110 мин. При подготовке корпуса каждый элемент произносился изолированно и сохранялся в отдельный файл. В табл. 2 приведены спектральные и временные характеристики акустических элементов в собранном корпусе. Использованы следующие сокращения: name — обозначение акустического элемента; count — количество элементов данного типа в корпусе; dur_mean — средняя длительность элемента; dur_v_mean — средняя длительность вокализованного участка элемента; dur_u_mean — средняя длительность невокализованного участка элемента; count_pitch — количество элементов данного типа с вокализованным участком в собранном корпусе; pitch_mean — средняя частота основного тона вокализованного участка элемента.

Анализируя временные характеристики элементов, следует отметить, что длительность записанных в корпусе звуковых файлов варьируется от 0,5 до 9 с. Однако неречевые элементы, связанные с деятельностью человека, в основном имеют длительность до 2 с. Звуки, не связанные с человеком, чаще всего имеют стационарный характер спектра, поэтому для обучения их моделей также требуется отрезок не более 2 с, несмотря на то, что в корпусе часть этих элементов имеет большую длительность, например, файлы со звуком течения воды имеют среднюю длительность 8,6 с.

■ ***Таблица 2.*** Спектральные и временные характеристики элементов в корпусе

name	count, шт.	dur_mean, мс	dur_u_mean,		dur_v_mean,		count_pitch, шт.	pitch_mean, Гц
			мс	%	мс	%		
th	200	1,691	1,634	96,6	0,056	3,4	138	388
mo	100	2,164	0,98	45,3	1,184	54,7	100	295
co	100	1,684	1,452	86,2	0,232	13,8	100	443
ap	200	2,831	2,823	99,7	0,007	0,3	6	517
fa	400	1,519	1,466	96,5	0,052	3,5	127	373
st	100	0,743	0,59	79,4	0,153	20,6	81	477
kd	200	1,602	1,555	97,1	0,047	2,9	100	478
kj	200	3,431	3,362	98,0	0,069	2,0	56	531
cm	200	2,434	1,979	81,3	0,455	18,7	200	282
ds	100	4,934	3,18	64,5	1,754	35,5	100	402
pw	200	3,399	3,352	98,6	0,046	1,4	58	478
pr	100	5,733	0,941	16,4	4,792	83,6	100	359
wa	100	8,671	8,158	94,1	0,513	5,9	100	525
ys	100	1,285	0,626	48,7	0,658	51,3	100	132
no	100	1,143	0,438	38,3	0,705	61,7	100	161
hl	100	1,786	0,699	39,1	1,087	60,9	100	152
tr	100	1,745	0,65	37,2	1,095	62,8	100	131

Что касается спектральных характеристик, то из табл. 2 видно, что большинство неречевых элементов в основном состоит из невокализованных участков. Исключение составляет телефонный звонок, однако, учитывая то, что мелодия звонка сейчас выбирается пользователем самостоятельно, параметрические характеристики могут здесь варьироваться в широком диапазоне. Соотношение вокализованных и невокализованных участков в речи примерно одинаковое. Приведенные характеристики были рассчитаны путем анализа аудиофайлов из собранного корпуса с помощью программы PRAAT [14].

Оценка работы системы распознавания акустических событий в ассистивном интеллектуальном пространстве производилась в два этапа. Вначале участник эксперимента осуществлял движение между заранее определенными точками, которые расположены под каждым из микрофонов и в центре помещения. При этом в каждой

точке движения акустическое событие моделировалось несколько раз. На втором этапе экспериментов участник свободно перемещался по комнате, моделируя различные события.

В ходе проведения экспериментов были записаны 2811 аудиофайлов. Анализ результатов распознавания показал, что большинство речевых команд распознавалось с точностью свыше 90 %. События «Падение» и «Звон ключей» были распознаны с меньшей точностью. Это связано со сложностью определения событий с низким уровнем энергии аудиосигнала, присутствующих в аудиосигнале вместе с сильным внешним шумом. Кроме того, при моделировании акустических событий существует возможность наложения одного события на другое, вследствие чего возникают ошибки. Средняя точность распознавания по всем элементам в корпусе составила 95 %, что вполне приемлемо для систем автоматической обработки акустической информации.

Заключение

Системы автоматической обработки акустической информации имеют важное значение во многих отраслях жизнедеятельности человека. Это связано с тем, что бесконтактный способ регистрации аудиоданных позволяет безопасно и ненавязчиво проанализировать текущую ситуацию в помещении и поведение человека. Средства аудиообработки сейчас успешно применяются в ассистивных системах наравне с методами обработки видео и данных, регистрируемых контактным путем. Представленная система распознавания акустических событий в интеллектуальном пространстве использована для определения потенциально опасных ситуаций, возникающих при повседневной жизнедеятельности. В ходе исследования собран акустический корпус для моделирования речевых и неречевых звуков, а также проведены эксперименты по их распознаванию. Последующая работа будет направлена на интеграцию разработанного метода с системой оповещения о потенциальных экстренных ситуациях в наблюдаемом помещении.

Работа выполнена в рамках федеральной целевой программы «Исследования и разработки по приоритетным направлениям развития научно-технологического комплекса России на 2007–2013 годы» (ГК № 11.519.11.4025).

Литература

1. **Portet F.** et al. Design and evaluation of a smart home voice interface for the elderly: acceptability and objection aspects // Personal and Ubiquitous Computing. 2011. Vol. 32. Iss. 1. P. 1–18.
2. **Chan M., Campo E., Esteve D., Fourniols J.** Smart Homes — Current Features and Future Perspectives // Maturitas. 2009. Vol. 64. Iss. 2. P. 90–97.
3. **Breazeal C.** et al. Humanoid robots as cooperative partners for people // Intern. J. of Humanoid Robots. 2004. Vol. 1. N 2. P. 1–34.
4. **Карпов А. А., Акарун Л., Ронжин Ал. Л.** Многомодальные ассистивные системы для интеллектуального жилого пространства // Тр. СПИИРАН. 2011. Вып. 19. С. 48–64.
5. **Popescu M., Li Y., Skubic M., Rantz M.** An acoustic fall detector system that uses sound height information to reduce the false alarm rate // Proc. of 30th Annual Intern. IEEE EMBS Conf., Aug. 2008. P. 4628–4631.
6. **Istrate D., Vacher M., Serignat J.** Embedded Implementation of Distress Situation Identification Through Sound Analysis // The J. on Information Technology in Healthcare. 2008. Vol. 6(3). P. 204–211.
7. **Chahuara P., Portet F., Vacher M.** Location of an inhabitant for domotic assistance through fusion of audio and non-visual data // Proc. of 5th Intern. Conf. on Pervasive Computing Technologies for Healthcare, May 2011. P. 242–245.
8. **Eronen A. J.** et al. Audio-based context recognition // IEEE Transactions on Audio, Speech, and Language Processing. 2006. Vol. 14. N 1. P. 321–329.
9. **Mesaros A., Heittola T., Eronen A., Virtanen T.** Acoustic event detection in real life recordings // Proc. of 18th European Signal Processing Conf., Aug. 2010. P. 1267–1271.
10. **Heittola T., Mesaros A., Eronen A., Virtanen T.** Audio context recognition using audio event histograms // Proc. of 18th European Signal Processing Conf., Aug. 2010. P. 1272–1276.
11. **Ma L., Milner B., Smith D.** Acoustic environment classification // ACM Trans. Speech Lang. Process. 2006. Vol. 3. N 2. P. 1–22.
12. **Ронжин Ал. Л., Ронжин Ан. Л.** Система аудиовизуального мониторинга участников совещания в интеллектуальном зале // Докл. ТУСУРа. 2011. № 1(22). Ч. 1. С. 153–157.
13. **Ронжин А. Л., Карпов А. А., Кагиров И. А.** Особенности дистанционной записи и обработки речи в автоматах самообслуживания // Информационно-управляющие системы. 2009. № 5. С. 32–38.
14. **Boersma P., Weenink D.** Praat: doing phonetics by computer. 2006. http://www.fon.hum.uva.nl/praat (дата обращения: 20.03.2012).