OSTIS-2012
(Open Semantic Technologies for Intelligent Systems)

УДК 004.934.2

СИСТЕМА ОПРЕДЕЛЕНИЯ ЭМОЦИОНАЛЬНОГО СОСТОЯНИЯ ДИКТОРА ПО ГОЛОСУ

Киселёв В.В., Давыдов А.Г., Ткаченя А.В.

ООО «Речевые технологии», г. Минск, Республика Беларусь
info@speetech.by

В статье рассматривается анализ речевой коммуникации в интеллектуальных диалоговых системах телекоммуникационной сферы. Речевая аналитика – новое направление в области речевых технологий ориентированное, на автоматический анализ разговора с целью выявления удовлетворенности собеседника. В статье кратко представлены теоретические аспекты системы определения эмоционального состояния диктора по голосу и практическая реализация предложенных методов на примере программного комплекса.
Ключевые слова: голосовой анализ, машинное обучение, распознавание эмоций, эмоциональное состояние диктора.

Введение

Первые попытки определения эмоциональных состояний с использованием статистических закономерностей, присущих определенным акустическим особенностям голоса, были предприняты в середине 80-х годов [Van Bezooijen, 1984] и [Tolkmitt and Scherer, 1986]. В дальнейшем интерес к задаче автоматического распознавания эмоций по голосу только возрастал. Количество ежегодно публикуемых статей, относящихся к данной тематике, с середины 90-х годов до настоящего момента, возросло более чем в 10 раз [Schuller et al.,2011], а эффективность распознавания приблизилась к человеческой.

Однако, несмотря на достигнутые успехи в сфере голосового детектирования эмоциональных состояний, проблема еще очень далека от своего окончательного решения. Существует ряд серьезных препятствий, значительно усложняющих разработку эффективных систем распознавания эмоционального состояния.

Прежде всего, нет четкого определения понятия «эмоция». При любых попытках его формализации мы натыкаемся на многообразие психологических моделей эмоциональных процессов, демонстрирующих различные ракурсы восприятия и модели описания эмоционального состояния диктора.

Вторая сложность – отсутствие единой теории, связывающей внутренние состояния диктора с особенностями его речи. Несмотря на достигнутые в этой сфере успехи, общепринятого подхода пока не существует [Scherer, 2003].

1. Система определения эмоционального состояния

Задача определения эмоционально состояния диктора широко востребована в современном мире, особенно в сфере оказания справочных услуг абонентам контакт-центров, для которых уровень удовлетворенности клиента оказанной ему услугой является главным критерием высокого качества работы.

Однако, в связи с большим количеством обрабатываемых запросов, невозможно производить ручную оценку удовлетворенности клиентов, отслеживать негативные звонки, «тихие» рекламации и пр. В этом случае для решения задачи определения эмоционального состояния диктора прибегают к использованию интеллектуальных автоматических систем распознавания эмоций.

В последние годы использование наукоемких алгоритмов голосового анализа, и в частности развитие методов распознавания эмоций, позволили получать оценку эмоций в реальном времени. Это позволяет использовать разрабатываемые системы не только для получения конечной оценки удовлетворенности абонента контакт-центра, но и для отслеживания изменения эмоционального состояния клиента во время его звонка.

Использование таких систем дает возможность четко контролировать работу операторов контакт-центров, с одной стороны своевременно сигнализируя им об малейших изменениях в эмоциональном состоянии абонента, позволяя

оператору быстро скорректировать свои дальнейшие действия, тем самым, предотвратив дальнейшее развитие конфликтной ситуации. А с другой стороны запись и анализ полученных эмоций могут быть использованы для выявления недобросовестности оператора контакт-центра.

Так же система определения эмоционального состояния может быть использованы в обучающих целях манеры ведения телефонных разговорах.

2. Определение информативных признаков

Важнейшим звеном системы автоматического детектирования эмоций по голосу диктора является выделение оптимального набора информативных признаков, коррелированных с эмоциональными состояниями. Выбор информативных признаков оказывает значительное влияние на эффективность классификации.

Акустические характеристики голоса могут быть условно разделены на пять категорий [El Ayadi et al., 2011]: просодические (частота основного тона, темп речи и т.д.), динамические (фонетическая функция), фонационные (отношение гармоник основного тона к шуму, джиттер, шиммер, и др.), спектральные (линейные спектральные частоты, кепстральные коэффициенты линейной шкалы частот, кепстральные коэффициенты мел-шкалы частот, и др.) и энергетические (отношение мощностей в спектральных полосах, оценка мощности сигнала и другие, как правило, основанные на энергетическом операторе Тигера). Каждая группа показателей предназначена для описания отдельных аспектов голоса, и находит свое применение в распознавании эмоциональных состояний.

Большинство исследователей уверены, что просодические характеристики речи, такие, как частота основного тона и энергия, передают значительную часть эмоционального содержимого высказывания [Cowie *et al.*, 2001], [Busso et al., 2009] и [Bosch, 2003]. Вильямс и Стивенс показали [Williams and Stevens, 1981], что степень возбуждения диктора влияет на суммарную энергию, распределение энергии по частотному спектру и на частоту и длительность пауз в речевом сигнале.

В качестве просодических характеристик можно выделить параметры, связанные с основным тоном, формантами, энергией, длительностями и артикуляцией [Cowie et al., 2001], [Murray and Arnott, 1993] и [Lee and Narayanan, 2005].

Значительные успехи в распознавании эмоций были достигнуты за счет генерации статичных векторов информативных признаков, получаемых из просодических характеристик сигнала после применения набора статистических функционалов. Предполагается, что данный факт косвенно свидетельствует о суперсегментной природе эмоциональной речи.

Джонстон и Шерер [Johnstone and Scherer, 2000] проанализировали результаты большого числа исследований, проведенных в этой области, и представили результаты в форме таблицы:

Таблица 1 – Акустические паттерны базовых эмоциональных состояний

	Стресс	Гнев
Интенсивность	↑*	↑
F0 уровень/среднее	↑	↑
F0 вариабельность		↑
F0 диапазон		↑
Контуры выражения		↓
Энергия ВЧ компонент		↑
Скорость речи и артикуляции		↑

продолжение Таблицы 1

Страх	Печаль	Радость	Скука
↑	↓*	↑	
↑	↓	↑	
	↓	↑	↓
↑(↓)	↓	↑	↓
	↓		
↑	↓	(↑)	
↑	↓	(↑)	↓

где символ «↑» обозначает увеличение параметра, символ «↓» – уменьшение.

3. Выбор наиболее информативных и помехоустойчивых параметров речи

Число всевозможных информативных признаков, выделяемых из звукового сигнала, может достигать нескольких тысяч. Далеко не все из них эффективны для решения задач распознавания эмоционального состояния, а немалая часть потенциально полезных характеристик оказывается избыточными. Перед построением и обучением классификаторов проводится предварительная процедура отбора информативных признаков. Конечной целью данного этапа работы является выделение релевантного набора характеристик звукового сигнала и декорреляция пространства информативных признаков.

При разработке современных систем распознавания эмоционального состояния диктора по голосу для минимизации набора информативных признаков широко применяются метод главных компонент и метод выделения признаков (feature selection). Их использование позволяет значительно сократить объем работы эксперта для построения наиболее эффективной системы. Вероятно, на данный момент наиболее популярной является стратегия линейного последовательного поиска (*sequential forward search)* [Pudil et al., 1994].

4. Алгоритмы классификации эмоционального состояния диктора

Системы автоматического распознавания эмоций состоят из двух основных блоков – первый осуществляет акустическую обработку входного речевого сигнала, выделяя из него набор

информативных признаков, отобранных в ходе предварительно проведенного исследования, а второй содержит классификатор, распознающий на их основе эмоциональное состояние диктора. Практически во всех предложенных системах распознавания эмоций были задействованы традиционные классификаторы, однако многие современные исследователи фокусируют свое внимание на разработке новых подходов.

Наиболее популярными техниками классификации являются следующие [Pantic and Rothkrantz, 2003]: поиск ближайших соседей (kNN), модель скрытых марковских процессов (HMM), модель смеси нормальных распределений (GMM), модели на основе нечеткой логики, искусственных нейронных сетей, байесовские классификаторы максимума вероятности, метод опорных векторов (SVM). На данный момент отсутствует какое-либо общепризнанное мнение по поводу того, какой классификатор лучше использовать.

При условии предварительно проведенного отбора информативных признаков, высокую эффективность дают такие классификаторы как линейный дискриминантный классификатор (LDC) либо поиск ближайших соседей (kNN), показавших свою эффективность как на идеальных [Dellaert et al., 1996] и [Petrushin, 1999], так и на более реалистичных базах [Batliner et al., 2000], [Kwon et al., 2003], [Lee and Narayanan, 2005] и [Shami and Verhelst, 2007].

5. Система автоматического определения эмоционального состояния

На основе вышеизложенной теории для разработки системы автоматического распознавания эмоций было принято решение использовать просодические характеристики речи в качестве информационных признаков эмоций, а классификацию эмоций проводить на основе поиска ближайших соседей (kNN).

Выбранная теория была реализована в программном комплексе, предназначенном для определения эмоционального состояния диктора по голосу.

В качестве просодических признаков эмоций были выбраны следующие характеристики голоса:

- интонированность: характеризует изменение производной частоты основного тона (ЧОТ). ЧОТ отражает высоту голоса диктора. Изменения (производная) ЧОТ определяют интонации голоса. Для монотонной речи характерны малые абсолютные значения производной ЧОТ (группировка значений около нуля) рисунок 1а. Излишне интонированная речь характеризуется значительным разбросом производной ЧОТ рисунок 1б;
- громкость – мощность сигнала;
- ритмичность – параметр, который отражает среднюю длительность фраз диктора;
- мелодичность – параметр, отражающий долю голосовых (вокализованных) фрагментов в речи;
- скорость – параметр, характеризующий темп речи, количество произносимых звуков в единицу времени.

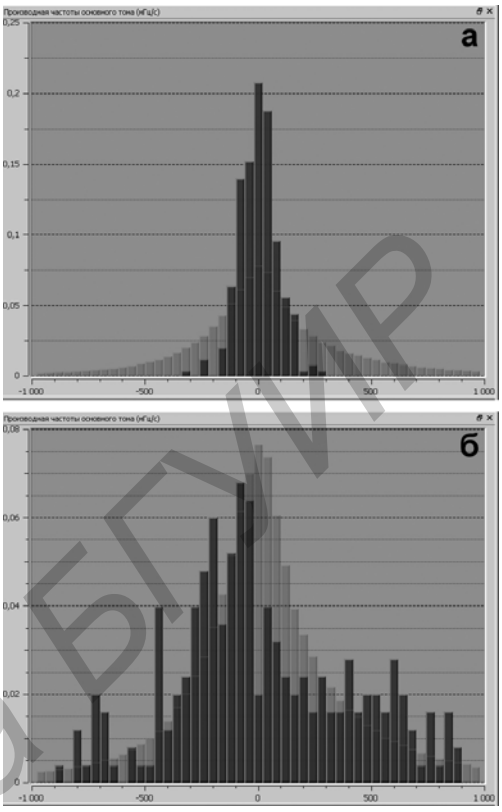

Рисунок 1 – Гистограмма интонированности речи (а – монотонная, б – излишне интонированная речь).

Гистограмма для достаточно выразительной речи показана зеленым цветом (более светлый оттенок)

Детектирование эмоций осуществляется по вхождению каждого из значений моментальных просодических характеристик голоса в соответствующие им диапазоны, заданные для каждой из эмоций.

В окне настройки просодических параметров эмоции также присутствуют параметры минимальной продолжительности эмоции и относительный коэффициент эмоции, который характеризует степень удовлетворенности (от 0 – полное отсутствие положительных эмоций, до 100 – максимально комфортное психоэмоциональное состояние). Оба этих параметра также учитываются при определении эмоции и вычислении общей оценки удовлетворенности диктора.

На рисунке 2, представлен фрагмент речи с распознанными эмоциональными состояниями.

Для увеличения достоверности принятия решения в программе используются модули определения пола и возраста диктора. Полученные параметры значений пола и возраста учитываются при оценке удовлетворенности диктора, так как

изменение этих параметров приводит с дисперсии величин просодических характеристик голоса.

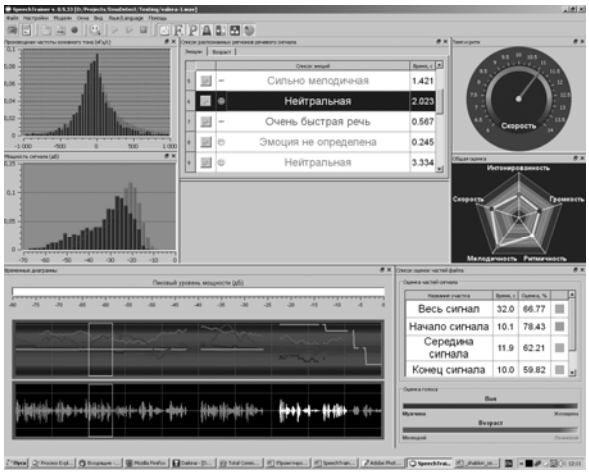

Рисунок 2 – Интерфейс программного комплекса

На ряду с определением эмоционального состояния диктора, программный комплекс позволяет просматривать текущие значения просодических характеристик голоса, создавать и настраивать просодические модели эмоций, а так же изменять параметры определения самих информативных просодических признаков. В связи с этим на базе программного комплекса можно проводить глубокий анализ просодических характеристик голоса.

Заключение

Речевые технологии предлагают пользователям широкий спектр автоматизированных услуг, одной из которых является автоматическая оценка эмоционального состояния диктора.

На основе изложенной в статье теории был разработан программный комплекс, который предназначен для повышения качества работы контакт-центра с клиентами при обслуживании обращений абонентов, нуждающихся в получении справочной информации. Преимущество использования разработанного ПО заключается в том, что при непрерывном притоке клиентов он обеспечивает постоянный контроль уровня качества обслуживания на любом этапе общения оператора с абонентом. Это позволяет центру обработки вызовов значительно повысить основные требования, предъявляемые к нему со стороны клиентов: скорость реакции на запросы абонентов, уровень удовлетворенности клиента и стремление к установлению эмоциональной связи с каждым абонентом.

Библиографический список

[Van Bezooijen, 1984] The Characterisitcs and Recognizability of Vocal Expression of Emotions / Van Bezooijen; Drodrecht // The Netherlands: Foris 1984.

[Tolkmitt and Scherer, 1986] Effect of experimentally induced stress on vocal parameters / Tolkmitt, F. J., Scherer, K. R. // J. Experimental Psychology: Human Perception and Performance 12(3), P. 302–313.

[Schuller et al.,2011] Schuller, B., Batliner, A., Steidl, S. and Seppi, D. Recognising realistic emotions and affect in speech // State of the art and lessons learnt from the first challenge. Speech Communication, In Press.

[Scherer, 2003] Vocal communication of emotion / Scherer, K.R. // A review of research paradigms. Speech Communication, 40(1-2), P. 227-256.

[El Ayadi et al., 2011] Survey on speech emotion recognition: Features, classification schemes, and databases / El Ayadi, M., Kamel, M.S. and Karray, F. // Pattern Recognition, 44(3), P. 572-587.

[Cowie et al., 2001] The description of naturally occurring emotional speech / Douglas-Cowie, E., Cowie, R. and Schroder, M. // Proc. 15th Internat. Conf. on Phonetic Sciences, Barcelona, Spain, P. 2877–2880.

[Busso et al., 2009] Analysis of emotionally salient aspects of fundamental frequency for emotion detection / Busso, C., Lee, S., Narayanan, S. // IEEE Trans. Audio Speech Language Process. 17(4), P. 582–596.

[Bosch, 2003] Emotions, speech and the asr framework / Bosch, L.; Speech Commun. 40, P. 213–225.

[Williams and Stevens, 1981] Vocal correlates of emotional states / Williams, C. and Stevens, K. // In. J. Darby (Ed.), Speech evaluation in psychiatry. New York: Grune & Stratton., P. 189–220.

[Murray and Arnott, 1993] Toward a simulation of emotions in synthetic speech: A review of the literature on human vocal emotion / Murray, I., Arnott, J. // J. Acoust. Soc. Am. 93(2), P. 1097–1108.

[Lee and Narayanan, 2005] Analysis of emotionally salient aspects of fundamental frequency for emotion detection / Busso, C., Lee, S., Narayanan, S. // IEEE Trans. Audio Speech Language Process. 17(4), P. 582–596.

[Johnstone and Scherer, 2000] Vocal communication of emotion / Johnstone, T., Scherer, K.R. // In: Lewis, M., Haviland, J. (Eds.), Handbook of emotion, second ed. Guilford, New York, P. 220–235.

[Pantic and Rothkrantz, 2003] Toward an Affect-Sensitive Multimodal Human–Computer Interaction / Pantic, M. and Rothkrantz, L.J.M., // Proceedings of the IEEE, 91(9), P. 1370-1390.

[Dellaert et al., 1996] Recognizing emotion in speech / Dellaert, F., Polzin, T., Waibel, A. // In: Proc. ICSLP, Philadelphia, PA, USA, P. 1970–1973.

[Petrushin, 1999] Emotion in speech: recognition and application to call centers / Petrushin, V. // In: Proc. Artificial Neural Networks in Engineering (ANNIE'99), St. Louis, MO, USA, P. 7–10.

[Batliner et al., 2000] Desperately seeking emotions: actors, wizards and human beings / Batliner, A., Fischer, K., Huber, R., Spiker, J. and Noth, E. // In: Proceedings of the ISCA Workshop Speech Emotion, P. 195–200.

[Kwon et al., 2003] Emotion recognition by speech signals / Kwon, O.-W., Chan, K., Hao, J., Lee, T.-W. // In: Proc. Interspeech, P. 125–128.

[Shami and Verhelst, 2007] An evaluation of the robustness of existing supervised machine learning approaches to the classification of emotions in speech / Shami M. and Verhelst W. // Speech communication, 49(3), P. 201-223.

[Pudil et al., 1994] Pudil, P., Novovicova, J., Kittler, J. (1994) Floating search methods in feature selection. Pattern Recognition Lett. 15, pp. 1119–1125.

SYSTEM OF DETERMINATION OF THE EMOTIONAL STATE OF THE SPEAKER OF A SOFTWARE TO THE VOICE

Kyselov V.V., Davydov A.G., Tkachenja A.V.

LLC "Speech Technology", Minsk, Belarus
info@speetech.by

The article describes the analysis of speech communication in intellectual dialogue systems of telecommunication. Speech analytics is a new direction in the field of speech technologies which is focused on the automatic analysis of conversation aimed at detecting the satisfaction of the interlocutor. The article briefly describes the theoretical aspects of the emotion recognition system and practical realization of the suggested methods through the example of bundled software.