Алгоритм эмоционального классификатора по речи для построения интерфейса человек-компьютер

Сидорова Ю.А.

МГУ им. М. В. Ломоносова

Введение

В рамках задачи оптимизации интерфейса человек-компьютер стоит задача обеспечения коммуникации между ЭВМ и человеком посредством голосовых команд. Интерес к распознаванию эмоций по голосу обусловлен тем, что существенная доля прагматически важной информации в речевой коммуникации передается невербально (в современной европейской культуре лингвосоциологические принципы коммуникации заставляют людей подавлять негативные эмоции и преувеличивать позитивные). Проблема автоматического распознавания эмоционального состояния говорящего по голосу на данный момент не является решенной.

Существующие системы различаются списками распознаваемых эмоций, типами используемых баз данных, акустическими параметрами и их производными а также алгоритмами классификаторов, эти различия делают результаты распознавания впрямую несопоставимыми. [Nogueiras et all, 2001] использовали скрытые модели Маркова и краткосрочные значения энергии и ЧОТ для распознавания между 7 эмоциональными состояниями на испанском языке с уровнем правильного распознавания, превышающим 80%. [Hozjan & Kacic, 2003] использовали нейронные сети и большой (144) набор статистических единиц, распознающих тех же самые 7 эмоциональных состояний на английском, словенском, испанском и французском, достигнув 60,33 - 88,73% правильного распознавания. [Toivanen et al, http] получили более чем 80% правильного распознавания, используя kNN классификатор с множеством из 41 статистических производных частично дикторонезависимого распознавания, распознавая 4 эмоции в финском языке.

Идеальный классификатор должен быть независим от диктора и естественного языка, работать в режиме реального времени, иметь высокую распознавательную мощность. Теоретическая возможность его построения на данный момент ни доказана, ни опровергнута.

Акустические параметры

В голосе эмоции «кодируются» акустическими параметрами речевого сигнала, основные из которых частота основного тона, интенсивность, темпоральные и спектральные характеристики. Дадим определения этих параметров. Частота основного тона (F0) - нижняя из частот сложной звуковой волны, слуховое впечатление, соответствующее параметру, – высота голоса. Интенсивность - средняя по времени энергия, которую звуковая волна переносит в единицу времени через единицу площади поверхности, расположенной перпендикулярно к направлению распространения волны; интенсивность звука пропорциональна квадрату амплитуды звукового давления; соответствующее слуховое впечатление – громкость. Из спектральных характеристик определим понятие форманты для гласных звуков, это область концентрации энергии в спектре звука, первые четыре форманты обозначаются соответственно F1, F2, F3, F4, слуховое впечатление – качество гласного (в смысле [а], [и], [е] и т. д).

В рамках данного эксперимента используются следующие акустические параметры: F0, интенсивность, темпоральные и спектральные характеристики. Вывод об эмоциональном состоянии говорящего делается на основе дескриптивной статистики этих

акустических параметров: математическое ожидание F0, медиана F0, максимум F0, дисперсия F0, среднее значение подъема огибающей F0, минимум значения подъема огибающей F0, максимум значения подъема огибающей F0, среднее значение подъема F0, минимальное значение падения F0, среднеквадратичное значение интенсивности, математическое ожидание среднеквадратической интенсивности, медиана среднеквадратической интенсивности, максимум среднеквадратической интенсивности, минимум среднеквадратической интенсивности, диапазон интенсивности, дисперсия интенсивности, средняя длительность гласного, средняя длительность согласных, темп, отношение речь/пауза, относительное количество энергии ниже 500 Гц, относительное количество энергии ниже 100 Гц, математическое ожидание F1, математическое ожидание F2, математическое ожидание F3, математическое ожидание F4, математическое ожидание ширины форманты F1, математическое ожидание ширины форманты F2, математическое ожидание ширины форманты F3, математическое ожидание ширины форманты F4.

Алгоритм эмоционального классификатора

Исследование исходит из гипотезы, что речевой сигнал информационно насыщен акустической информацией, и что эта акустическая информация отражает лингвистическую структуру [Hawkins et al, 1998]. Древовидные структуры используются для представления иерархичного устройства высказывания, их удобно описывать древесными автоматами. Напомним, что детерминированный древесный автомат – это четверка $A = (Q, V, \delta, F)$, где Q – конечное множество состояний, V – ранжированный алфавит, $Q \ni V = \downarrow$, $F \subseteq Q$ – множество конечных состояний, и $\delta = (\delta_0, .. , \delta_m)$ – конечное множество продукций, таких, что $\delta_n : (V_n \times (Q (V_0)^n) \rightarrow Q$, $n = 1, \ldots m$, $\delta_0 (a) = a$, $\forall a \subseteq V_0$.

Применен метод [Sempere, Lopez, 2003], состоящий их 2-х частей: (1) грамматический вывод для получения характеризации каждого типа эмоционального высказывания (гневного, радостного и т. д.); (2) получены расстояния между высказываниями и древесными автоматами с целью построить дерево принятия решений.

На этапе (1) система учит множество древесных автоматов (один на эмоцию), используя грамматический вывод (грамматический вывод – это задача нахождения множества правил неизвестной грамматики G из множества примеров, которые могут принадлежать иди не принадлежать L(G)).

На этапе (2) использован алгоритм машинного обучения C 4.5 [Quinlan, 1993]: вычисляется энтропия каждого узла древовидной структуры (соответствует акустическому параметру), по убыванию энтропии узлы располагаются на дереве принятия решения, то есть в корне дерева – самый мощный классифицирующий параметр, далее другие по убыванию, параметры, не влияющие на результат распознавания на пути решения не встречаются, «ветки» заканчиваются решением («гнев», «нейтральное», «радость» и т. д.).

Эксперимент

Языковой материал - фрагменты звукоряда фильма «Преступление и наказание» на русском, соответствующие речи четырех героев (324 высказывания).

Множество высказываний было разделено на два подмножества одинаковой мощности: одно, чтобы выучить классификатор (дерево принятия решений), а другое, чтобы протестировать его. Уровень правильного распознавания около 80% (под «правильным» распознаванием понимается автоматическое решение, совпадающее с оценкой, по меньшей мере, 10 из 14 носителей русского языка, производивших слуховую оценку). Результаты говорят в пользу того, что использованный метод может быть применен с целью автоматического распознавания эмоциональных состояний.

Таблица 1: матрица ошибок распознавания.

	Гнев	Ирония	Удив.	Страх	Неудов.	Радос.	Стыд	Нейтр.	Сумма%
Гнев	88%	0%	5%	0%	0%	2%	0%	5%	100%
Ирония	0%	75%	3%	0%	1%	1%	0%	10%	100%
Удивление	3%	0%	85%	3%	4%	3%	0%	2%	100%
Страх	0%	0%	21%	60%	3%	1%	0%	15%	100%
Неудов.	0%	0%	0%	0%	88%	0%	0%	12%	100%
Радос.	13%	0%	16%	4%	0%	67%	0%	0%	100%
Стыд	0%	0%	0%	4%	6%	0%	80%	10%	100%
Нейтрал.	1%	1%	10%	1%	8%	1%	1%	77%	100%

Выводы

Была предпринята попытка дикторозависимого распознавания с использованием ранее не применявшегося с этими целями метода, соединяющего грамматический вывод на регулярных древесных языках и деревья принятия решений. Уровень правильного распознавания – примерно 80%, таким образом метод применим для задачи распознавания эмоционального состояния говорящего по речи.

Библиография

[Nogueiras et all, 2001] Noguerias A., Moreno A., Bonafonte A., Marino J. B., “Speech emotion Recognition Using Hidden Markov Models”, Eurospeech 2001, 2001.

[Hozjan, Kacic, 2003] Hozjan V., Zdravko K. “Improved Emotion recognition with Large Set of Statistical Features”, Eurospeech 2003, 2003.

[Toivanen et al., http] Toivanen J., Seppanen T., Vayrynen E. Automatic recognition of emotions in spoken Finnish : preliminary results and applications, http://www.mediateam.oulu.fi/publications/pdf/404.pdf

[Hawkins et al, 1998] Hawkins, S., House, J., Huckvale M., Local J., Ogden R. “ProSynth: An Integrated Prosodic Approach to Device-Independent, Natural sounding Speech Synthesis”, International Conference Speech and Language Processing, 1998.

[Sempere, Lopez, 2003] Sempere J. M., Lopez D. “Learning decision trees and tree automata for a syntactic pattern recognition task”, 1st Iberian Conference on Pattern recognition and Image Analysis, 2003.

[Sakakibara, 1997] Sakakibara Y., “Recent advances of grammatical inference”, Theoretical Computer Science 185, pp 15-45, 1997.

[Quinlan, 1993] Quinlan J. R. C4.5: programs for machine learning. Morgan Kaufmann, 1993.