ММРО
математические методы распознавания образов
19-я Всероссийская конференция с международным участием
«Математические методы распознавания образов»
г. Москва, 26 - 29 ноября 2019 года

Выделение предварительно записанных голосовых сообщений в аудиозаписях телефонных разговоров

Копылов А.В.[1], Середин О.С.[1], Тышкевич Б.В.[2], Филин А.И.[12]

Тула, Тульский государственный университет[1]

Тула, IТooLabs Soft[2]

And.Kopylov@gmail.com, oseredin@yandex.ru, bvt@itoolabs.com, afilin@itoolabs.com

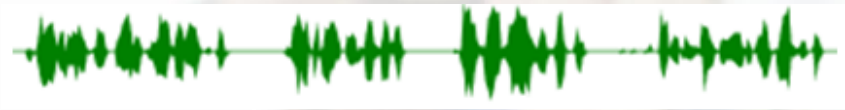
IToolabs
Российский разработчик ПО для операторов связи

https://itoolabs.com/ru/

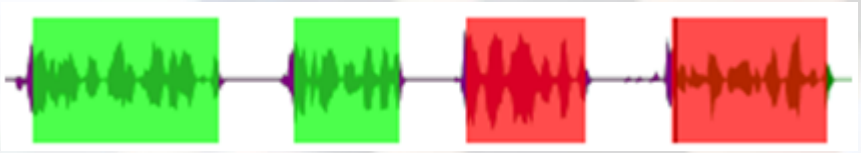
Тел.: +7 (495) 669-72-21
E-mail: hi@itoolabs.com
Адрес: Москва, Научный проезд д.8 с.1

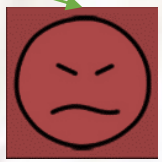
Оценка эмоционального фона диалога оператора и клиента

1. Запись диалога
2. Разбиение на фрагменты
3. Формирование признаков фрагментов
4. Классификация эмоций на фрагментах
5. Оценка эмоциональности записи

1.


2.


4.




Выделение предварительно записанных сообщений

Задача состоит в выделении фрагментов аудиозаписи, содержащих IVR, и удалении их из дальнейшего анализа эмоционального фона.

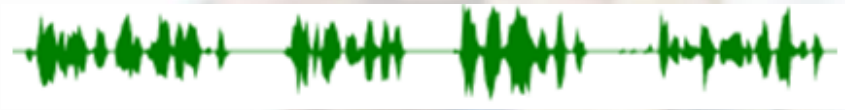
IVR (Interactive Voice Response) являются речевыми фрагментами, но произнесены диктором заранее, так же могут сопровождаться фоновой музыкой или включают чисто музыкальные фрагменты и воспроизводятся с целью автоматизации взаимодействия с клиентом (маршрутизация звонков, создание интерактивной очереди, «холодный» обзвон и т.п.)

Фильтрация IVR выходит за рамки традиционного выделения речевых фрагментов (Voice Activity Detection, VAD).

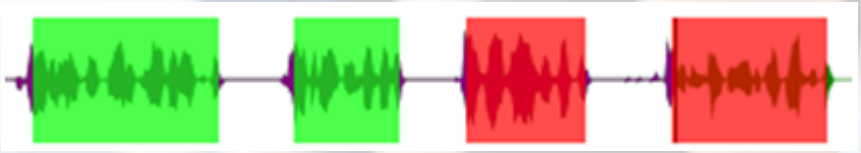
Оценка эмоционального фона диалога оператора и клиента

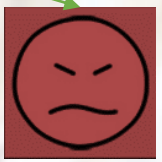
1. Запись диалога
2. Разбиение на фрагменты
3. Формирование признаков фрагментов
4. Классификация эмоций на фрагментах
5. Оценка эмоциональности записи

IVR

1.


2.


4.




База данных IТooLabs

Initial Class Labels	Audio Fragments	Class Labels	Audio Fragments
Happy	5569	Live Speech	179417
Neutral	138848		
Resentment	6782		
Vexation	1645		
Wrath	407		
Emotions mixed	2411		
Speech mixed	23755		
IVR	15660	IVR	38143
IVR PBX	5707		
IVR music or noise	5221		
IVR music mixed	27		
Music or noise	11528	Nonspeech	

Копылов А.В. et al. Формирование базы данных для системы оценки эмоционального фона диалога с оператором центра обработки вызовов // Всероссийская конференция ММРО-2017. Москва: ТОРУС ПРЕСС, 2017. Р. 132–133

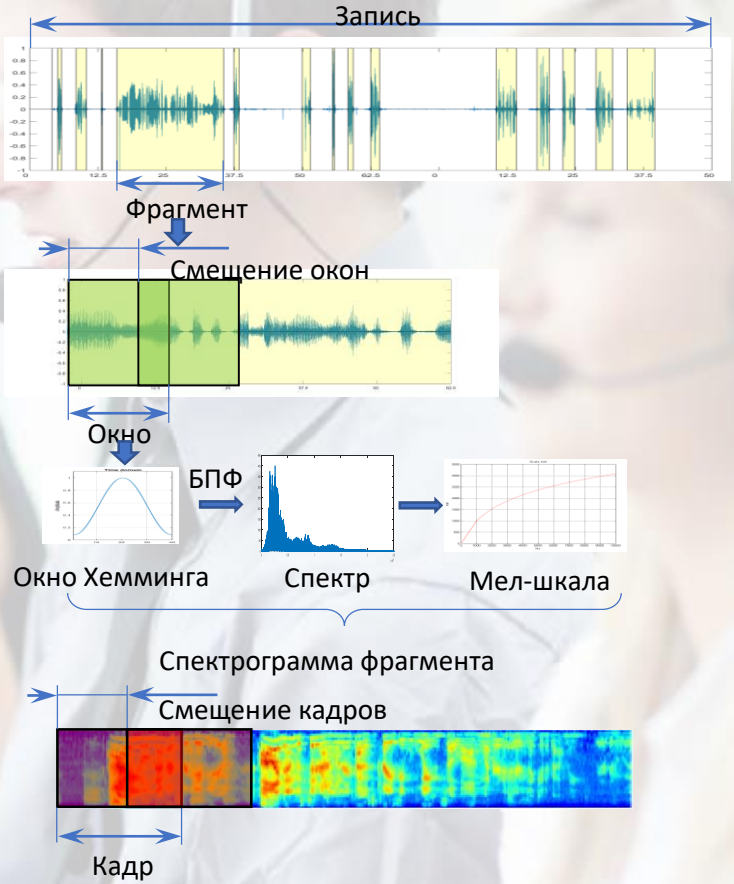
Существующие методы выделения IVR

- На основе распознавания тоновых сигналов, отмечающих начало или конец записи[1]
- На основе распознавания речи и статистики использования определенных фраз в ответе человека и машины[2, 3]
- Активное взаимодействие и/или определение резких изменений характеристик канала связи[4]

1. Voicemail detection powered by AI [Electronic resource]. URL: https://voximplant.com/blog/voicemail-detection-powered-by-ai (accessed: 22.10.2019).
2. Ju Y.C., Wang Y.Y., Acero A. Call analysis with classification using speech and non-speech features // INTERSPEECH 2006 9th Int. Conf. Spok. Lang. Process. INTERSPEECH 2006 - ICSLP. 2006. Vol. 4. P. 1902–1905.
3. Acero A. et al. Detecting an answering machine using speech recognition: pat. 8065146 USA. USA: Google Patents, 2011.
4. Shiota S. et al. Voice liveness detection algorithms based on pop noise caused by human breath for automatic speaker verification // Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH. 2015. Vol. 2015-Janua. P. 239–243.

Признаковое описание фрагментов аудиозаписей

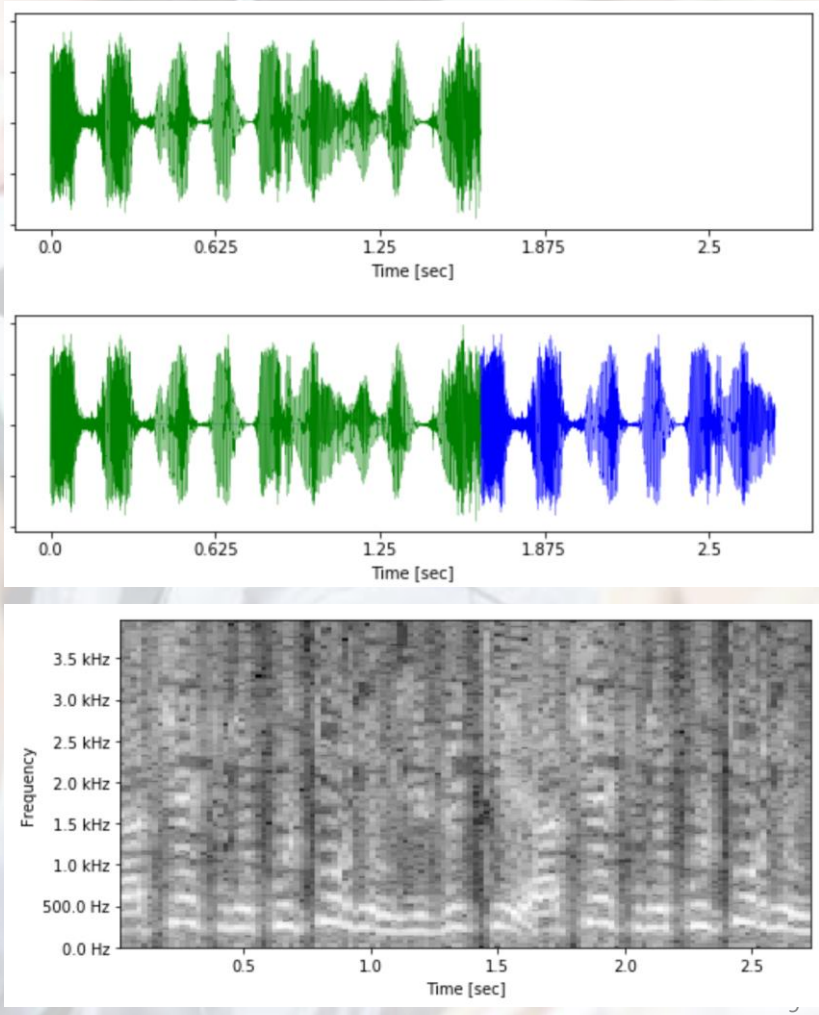
- Вычисляемые акустические признаки (набор GeMAPS[1]) – объектом распознавания является фрагмент
- Непосредственное представление записи (спектрограммы, гаммотонограммы) - объектом распознавания является кадр



1. Eyben F. et al. The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for Voice Research and Affective Computing // IEEE Trans. Affect. Comput. 2016. Vol. 7, № 2. P. 190–202.

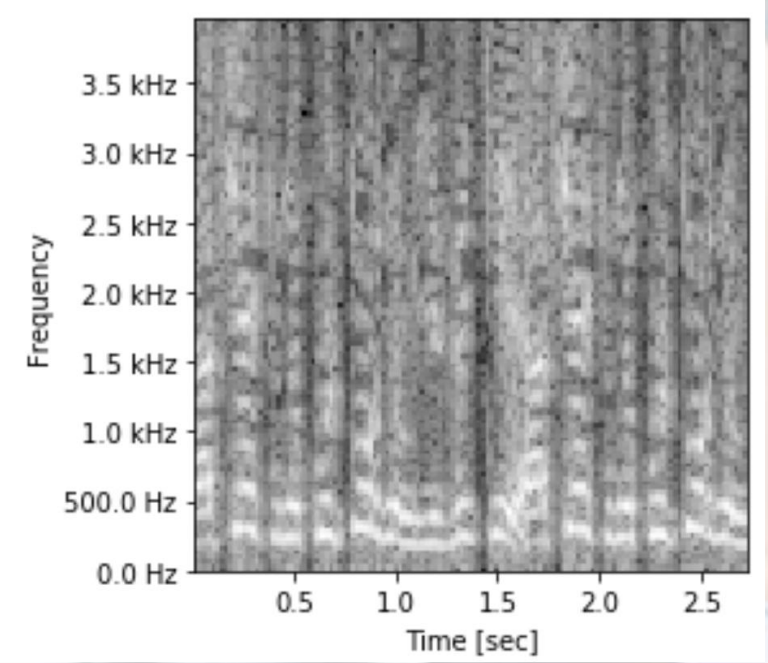
Приведение спектрограмм к одинаковой размерности

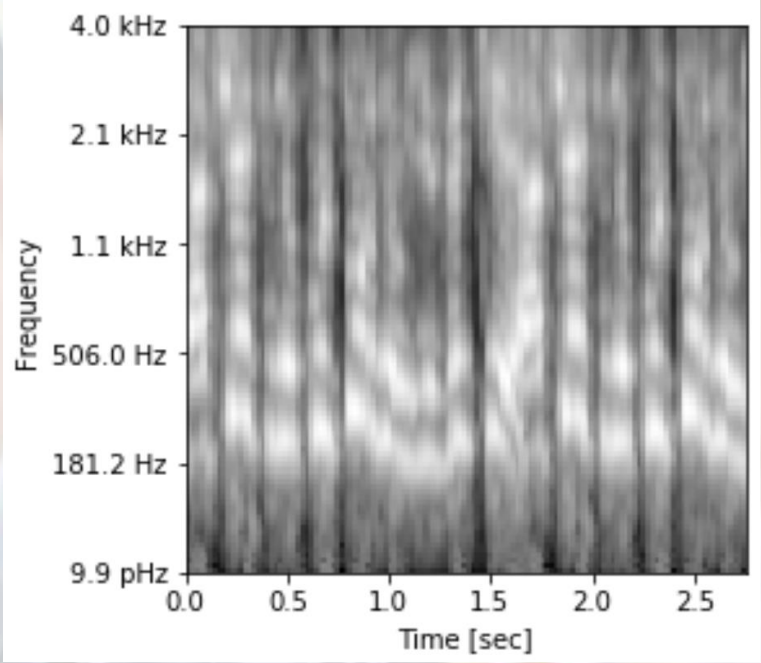
- t – длительность исходного фрагмента
- T = 2.7466 с – длительность фрагмента, к которой приводятся исходные фрагменты
- При t < T – *фрагменты* циклически добавляются до достижения длительности T
- При *T < t < 2T – фрагменты* обрезаются справа
- При t >= 2T – фрагмент разделяется на n неперекрывающихся фрагментов длительностью T



Приведение спектрограмм к одинаковой размерности

- Спектрограммы - кратковременное преобразование Фурье, ширина окна 512 отсчетов и перекрытием в 170 отсчетов с использованием весового окна Хемминга
- Гамматонограммы - 128-канальный банк фильтров





Методы классификации аудиофрагментов при наличии речевых IVR

- Метод опорных векторов (SVM[1])
- Градиентный бустинг (XGBoost[2])
- Сверточная нейронная сеть (CNN[3])

1. Vapnik V.N. Statistical Learning Theory // Interpreting / ed. Haykin S. Wiley-Interscience, 1998. Vol. 2, № 4. 736 p.
2. Riedman J. Greedy Function Approximation: A Gradient Boosting Machine // Ann. Stat. 2001. Vol. 29, № 5. P. 1189–1232.
3. Mu Y. et al. Speech Emotion Recognition Using Convolutional- Recurrent Neural Networks with Attention Model // Inf. Sci. Internet Technol. 2017. № Cii. P. 341–350.

Структура примененной CNN

Layer	Layer Parameters	Output Shape	Number of Parameters
Input		(128, 128, 1)	
Conv2d	(16, 5x5)	(122, 122, 16)	800
Batch_normalization		(122, 122, 16)	64
Activation	RELU	(122, 122, 16)	0
Max_pooling2d	(2x2)	(61, 61, 16)	0
Conv2d	(16, 3x3)	(57, 57, 16)	6416
Batch_normalization		(57, 57, 16)	64
Activation	RELU	(57, 57, 16)	0
Max_pooling2d	(2x2)	(28, 28, 16)	0
Conv2d	(16, 3x8)	(26, 26, 16)	2320
Batch_normalization		(26, 26, 16)	64
Activation	RELU	(26, 26, 16)	0
Max_pooling2d	(2x2)	(13, 13, 16)	0
Flatten		(2704)	0
Dense		(676)	1828580
Batch_normalization		(676)	2704
Activation	RELU	(676)	0
Dense		(128)	86656
Batch_normalization		(128)	512
Activation	RELU	(128)	0
Dense		(32)	4128
Batch_normalization		(32)	128
Activation	SOFTMAX	(32)	0
Dense		(2)	66

База данных IТooLabs

Initial Class Labels	Audio Fragments	Class Labels	Audio Fragments
Happy	5569	Live Speech	179417
Neutral	138848		
Resentment	6782		
Vexation	1645		
Wrath	407		
Emotions mixed	2411		
Speech mixed	23755		
IVR	15660	IVR	38143
IVR PBX	5707		
IVR music or noise	5221		
IVR music mixed	27		
Music or noise	11528	Nonspeech	

Копылов А.В. et al. Формирование базы данных для системы оценки эмоционального фона диалога с оператором центра обработки вызовов // Всероссийская конференция ММРО-2017. Москва: ТОРУС ПРЕСС, 2017. Р. 132–133

Результаты экспериментов для 10-кратной кросс-проверки

GeMAPS + SVM		
True / Predicted	Live Speech	IVR +Nonspeech
Live Speech	0.9904	0.0096
IVR+Nonspeech	0.1339	0.8661
Macro Avg F1		0.9437

GeMAPS + XGBoost		
True / Predicted	Live Speech	IVR +Nonspeech
Live Speech	0.9929	0.0071
IVR+Nonspeech	0.1753	0.8247
Macro Avg F1		0.9329

Logspectra + CNN		
True / Predicted	Live Speech	Non speech
Live Speech	0.9837	0.0163
IVR+Nonspeech	0.0283	0.9717
Macro Avg F1		0.9778

Gammatone filterbank + CNN		
True / Predicted	Live Speech	IVR +Nonspeech
Live Speech	0.9923	0.0077
IVR+Nonspeech	0.0318	0.9682
Macro Avg F1		*0.9821*

Спасибо за внимание!