



Engaging Content
Engaging People

Unsupervised Clustering for Expressive Speech Synthesis

João P. Cabral
Trinity College Dublin, Ireland

Machine Learning Meetup
28 May 2018

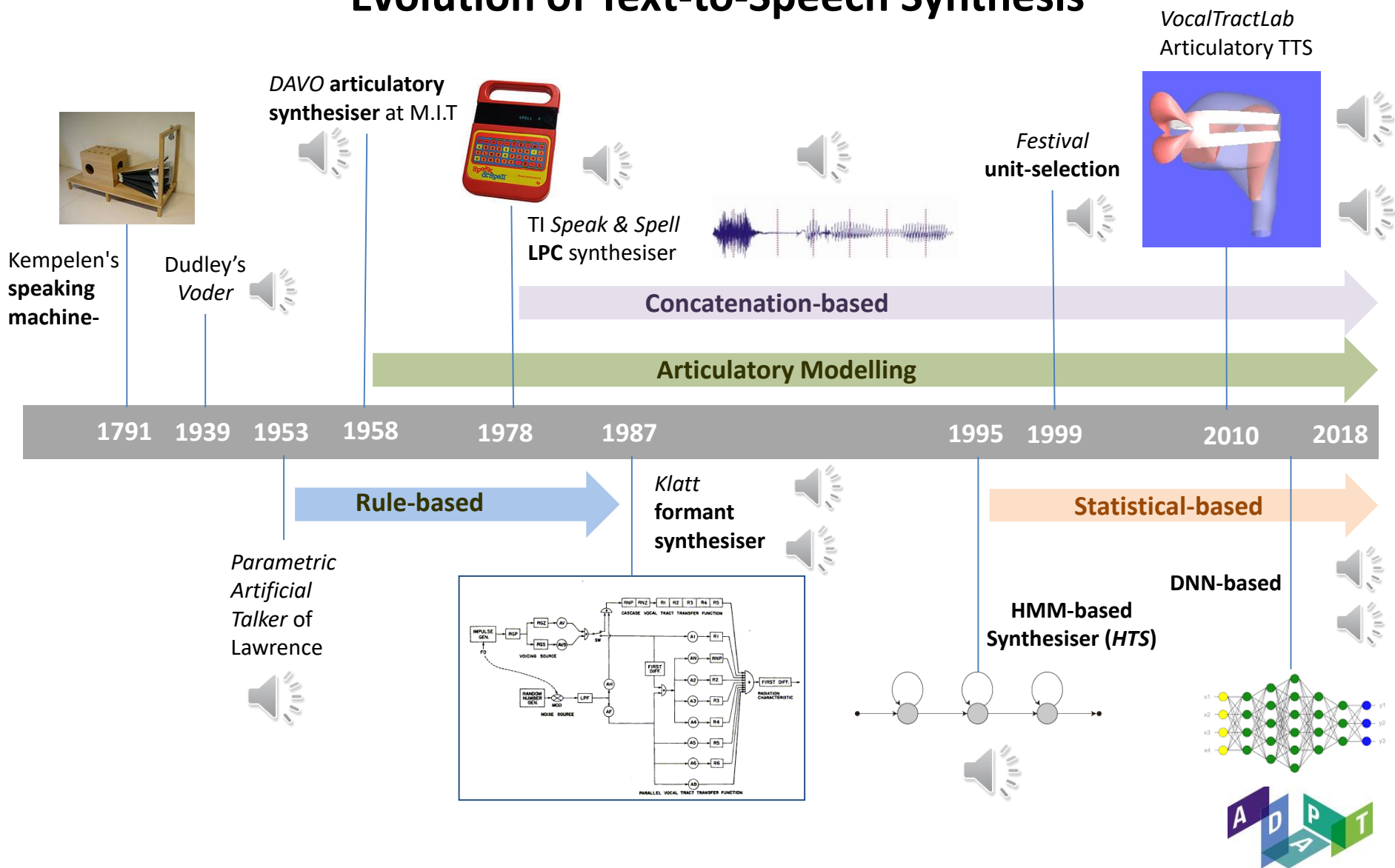


The ADAPT Centre is funded under the SFI Research Centres Programme (Grant 13/RC/2106) and is co-funded under the European Regional Development Fund.

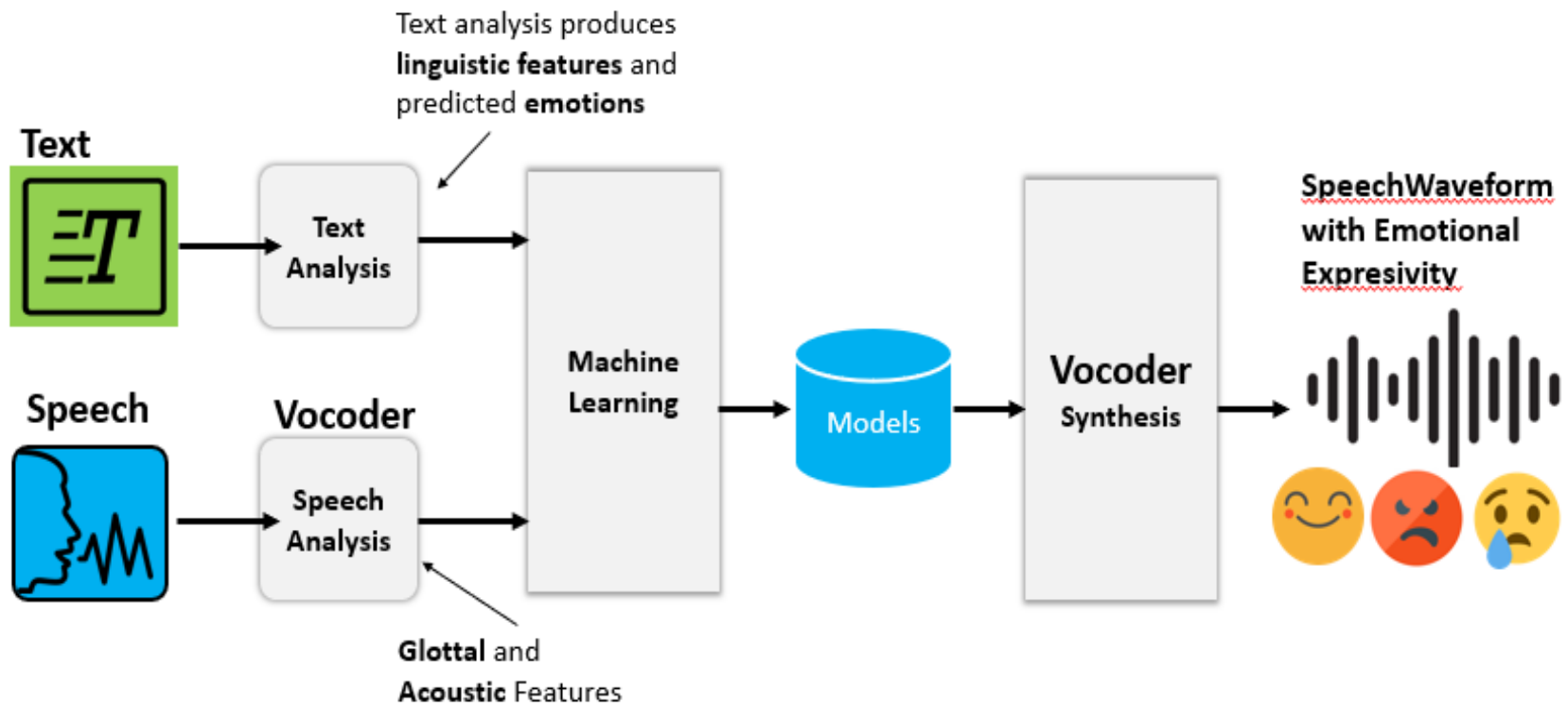
- Introduction to Text-to-Speech Synthesis
- Emotion Classification from Text
- Semi-automatic Emotion Labeling using Unsupervised Speech Clustering
- Summary and Future Directions



Evolution of Text-to-Speech Synthesis



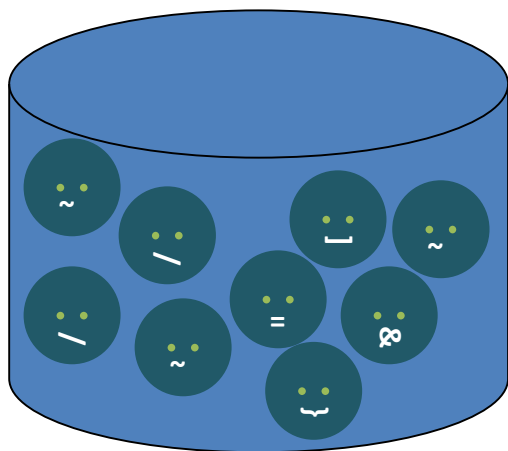
Expressive TTS System with glottal features



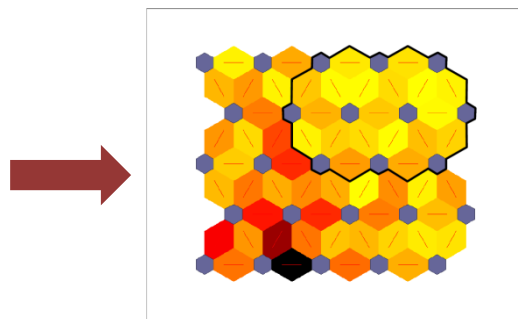
(Cabral et al., 2011)

Clustering of Voice Styles

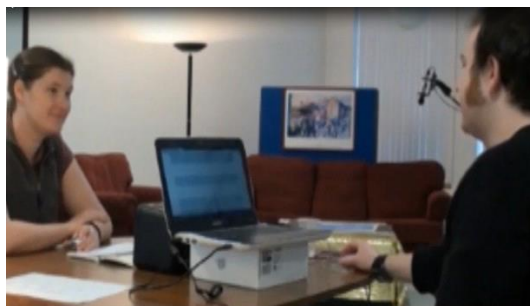
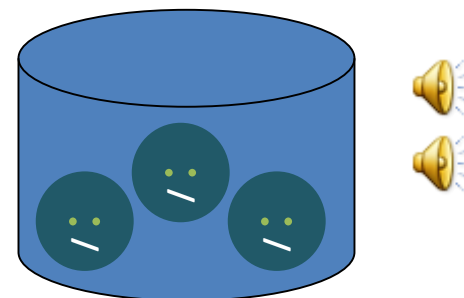
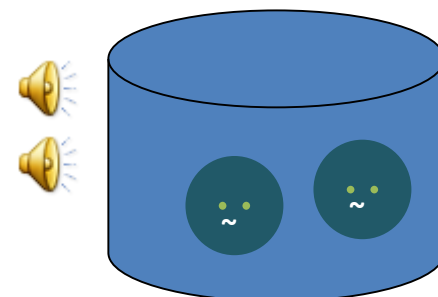
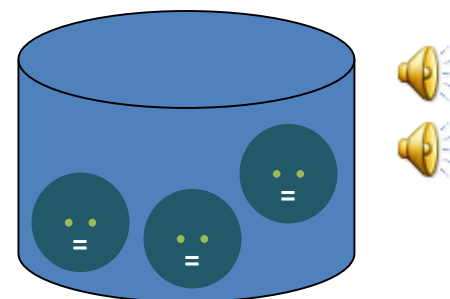
Speech from audiobook with
different voice styles



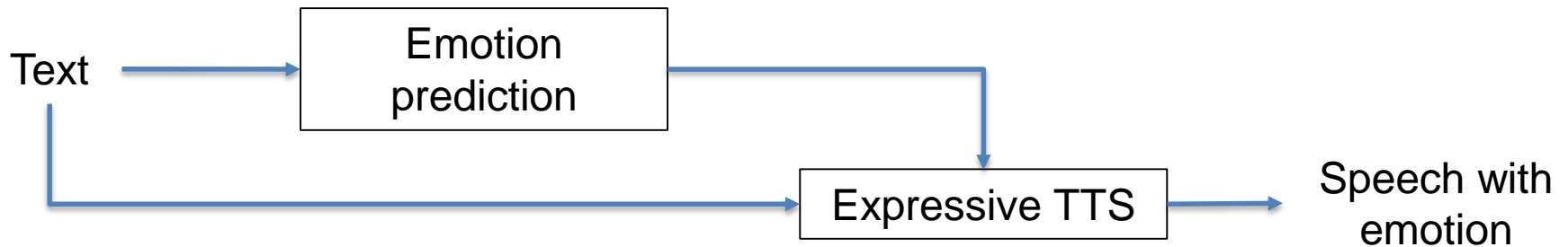
Clustering using the
glottal source parameters



Subsets of associated with
different voice styles



Synthesis of Emotional Speech

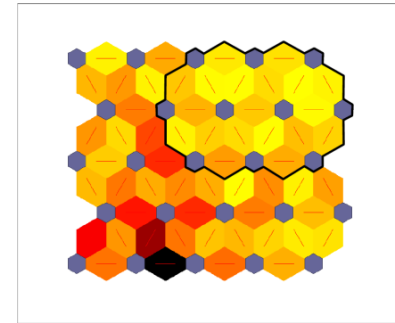


- Emotion Prediction from Text:
 - Machine Learning vs Dictionary-based
 - Problem with inter-speaker and intra-speaker variability
- Modeling and Generating Emotional Speech
 - Lack of audiobook corpora with annotations of emotions
 - Emotion perceived from speech and text may be different
 - Acoustic correlates of emotion are still not well known

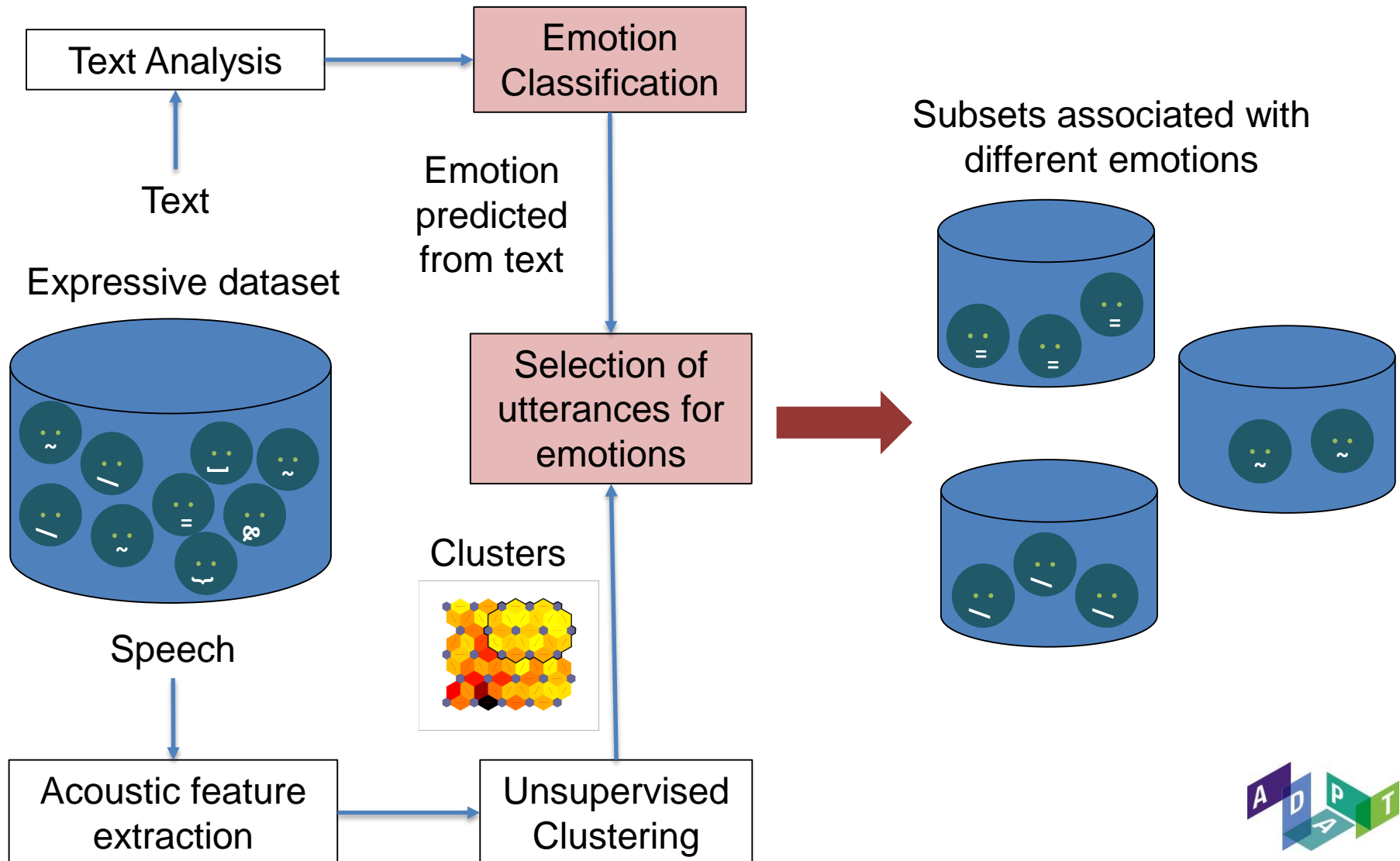
- Motivation: Improve expressiveness of synthetic speech and convey essential non-linguistic information
- Focus on storytelling of fairy tales and emotions:
 - Expressivity and emotions are fundamental for reader engagement
 - Publicly available data resources
 - Speech emotions from narrative style, in contrast to recordings of isolated sentences with acted emotions
- Goals:
 - Predict emotions from text correctly
 - Alleviate manual labeling in training of expressive voices in statistical parametric speech synthesis



- Traditional approach using labeled speech emotions
 - Several training techniques: Decision Trees, Acoustic Model Adaptation, etc.
 - Problem with data preparation (time consuming and expensive)
 - Limitation with inter-speaker variation and limited emotions
- ~~Semi-automatic emotion labeling using speech clustering~~
 - Clusters may contain different emotions
 - Needs perceptual verification of expressions
 - Limitations for many speakers and emotions
- Joint training of linguistic and expressive acoustic spaces
 - Depends on correlation between acoustic and linguistic space
e.g. (Cheng, L. et al. 2013)
 - Large number of expressions and avoids inter-speaker factors
 - Limited control over synthesized emotions



Emotion Prediction from Text to Alleviate Annotation



Classification of Emotion Polarity:

- Classification into 3 categories: Positive, Negative, Neutral
- Use of the tool SentiWordsTweets
- ✓ High accuracy
- Not fine-grained

Classification of Emotion Category:

- Emotion of phrase is predicted using lexicon-based method by selecting top rated emotion in sentence
- NRC Emotion Lexicon (8 emotions) and vocabulary from fairy tales
- 7 basic emotions: Angry, Sadness, Surprise, Joy, Fear, Disgust, Neutral
- ✓ Fine grained to multiple number of emotions
- Low accuracy



Combination of Emotion Category and Polarity:

- A sentence is labeled into a specific category if the sentiment-polarity matches the polarity of the emotion
- Method to avoid 'over-tagging' of sentences with emotions:
 1. The highest count of emotions is divided by the number of tokens
 2. Select top rated emotion, above a given threshold
 3. Threshold can be derived from the human annotations of emotion

Examples of Emotion Prediction:

Sentences	<i>Sentiment-polarity</i> (range 0 to 1)	<i>Polarity of Emotion</i> <i>Category</i>	<i>Final Emotion Label</i>
<i>Juliet's dead</i>	Negative (score =0.35)	Negative (Fear)	Fear
<i>I mean lovely</i>	Positive (score=0.59)	Positive (Joy)	Joy
<i>What name did they give the child?</i>	Negative (score=0.44)	Positive (Joy)	<i>Neutral</i>

Evaluation:

- Corpus of emotion annotations (2 annotators): 176 fairy tales from Grimm (80 tales), H.C. Andersen (77 tales) and Potter (19 tales)
- Grimm's tales used for testing
- Extended NRC Emotion lexicon with vocabulary of the fairy tales

Results:

Emotions	Anger	Sadness	Joy	Fear	Surprise	Disgust	Neutral
Rate of Labels System	2.3%	2.8%	7.4%	1.9%	3.2%	0.8%	81.6%
Rate of Labels Annotators	4.1%	3.4%	6.2%	2.9%	1.6%	0.3%	81.5%
F-scores System	0.2	0.3	0.38	0.18	0.09	0	0.86
F-scores Annotator 1	0.41	0.39	0.53	0.32	0.38	0.09	0.71

Audiobook Corpus:

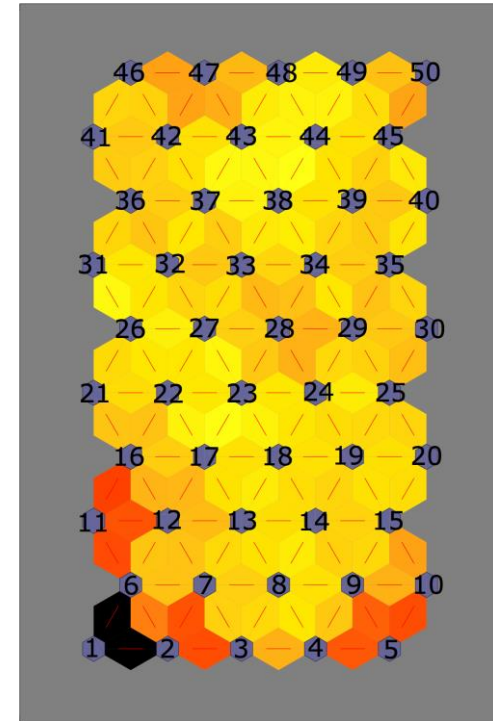
- Corpus of audiobooks released for the Speech Synthesis Blizzard Challenge 2016
- Performed sentence-level alignments between speech and text using Kaldi
- Selected emotional utterances from direct speech

Clustering Speech Styles:

- Self Organising Map (SOM)
- 605 acoustic features extracted with openSMILE
- Number of clusters was 50 based on informal listening tests

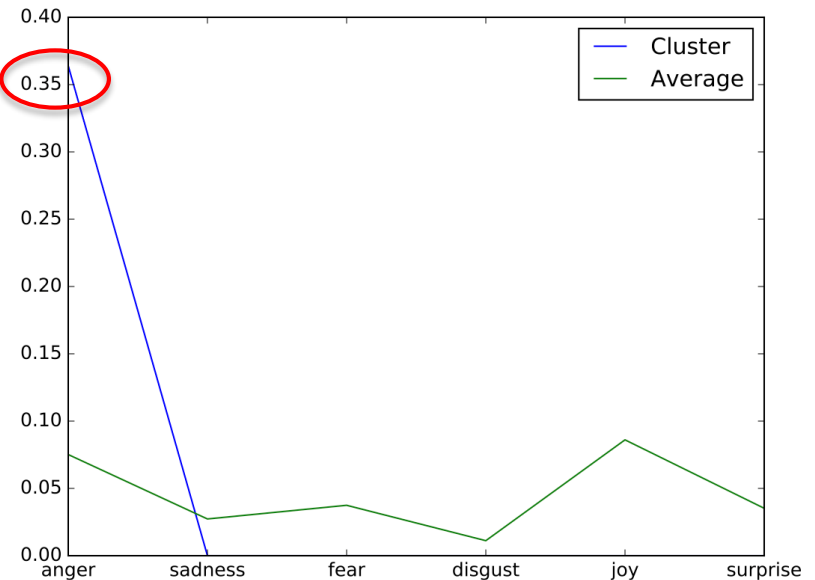
Mapping of Speech Clusters to Emotions:

- Performed emotion prediction from text on the sentences belonging to a specific cluster
- Compared rate of detected emotions in each cluster with the overall distribution of emotions over all the clusters
- Select candidate clusters for each emotion based on distance between clusters and results of emotion prediction



Distribution of Predicted Emotions (Cluster 11)

Expect that this cluster has a higher number of utterances with speech emotion “Anger”



Using Sentiment Analysis to Make Emotion Classification More Restrictive

- Decrease the threshold applied to the sentiment-polarity values in order to select sentences with stronger ‘sentiment level’ for an emotion
- For example, with threshold lower than 0.35 gives 10 ‘strongly negative’ sad sentences
- Classification of emotions of the utterances solely on prediction from text

Results of Automatic Emotion Classification:

- 50 random utterances labelled by the tool
- Author evaluated emotion labels by considering text without any context
- Author evaluated speech emotion by listening to utterances
- Calculated the rate of emotions perceived by the listener that matched the emotion predicted from text

























Emotion Labels	Text Emotion Classification	Speech Emotion Classification from Text	Match between Perceived Emotion and Correct Text Labels
Anger	63%	50%	79%
Sadness	78%	42%	54%
Joy	76%	51%	67%
Fear	56%	34%	61%
Surprise	74%	68%	91%
Disgust	33%	33%	100%
Average	63%	46%	75%

- Strong correlation between the emotion prediction from text and the corresponding speech emotion
- Automatic prediction tool can be useful in selection of speech with emotions



Synthesis of Speech with Emotions:

- Selected at least 20 utterances for each emotion (ranged from 26 to 54)
- HTS-2.3 system using MLLR adaptation
- STRAIGHT vocoder, with F0 calculated with RAPT algorithm
- Festival for text analysis

Emotions	Speech Synthesized with Emotion	Speech Synthesized with Neutral Voice
Joy	 	 
Fear	 	 
Surprise	 	 
Anger	 	 
Sad	 	 
Disgust	 	 

Concluding Remarks:

- Method that combines information of lexicon-based sentiment analysis with sentiment-polarity scores to improve accuracy of emotion labelling system
- Control over the number of sentences labelled with emotion by using threshold of sentiment polarity score
- Emotion predictions from text were close to those obtained by human annotation, indicating that some emotions are more difficult to predict (disgust, surprise, and fear)
- Emotion prediction tool can be helpful in the selection of subsets of audiobook data for building synthetic voices
- Integration of emotion prediction into HMM-based speech synthesizer

Future Directions:

- Conduct more extensive perceptual experiment to evaluate emotions of synthetic voice and correlation between emotions predicted from text and those conveyed in uttered speech
- Compare sentiment analysis method to other approaches, in particular non-dictionary based methods
- Integration into more advanced TTS systems



João P. Cabral, Steve Renals, Junichi Yamagishi and Korin Richmond, “HMM-based speech synthesiser using the LF-model of the glottal source,” International Conference on Acoustics, Speech and Signal Processing (ICASSP), Prague, Czech Republic, pp.4704-4707, 2011.

Langzhou Chen, Mark J. F. Gales, Norbert Braunschweiler, Masami Akamine and Kate Knill, “Integrated Expression Prediction and Speech Synthesis From Text,” IEEE Journal of Selected Topics in Signal Proc., 8(2):323-335, 2014.

Eva Vanmassenhove, João P. Cabral and Fasih Haider, “Prediction of Emotions from Text using Sentiment Analysis for Expressive Speech Synthesis,” 9th ISCA Workshop on Speech Synthesis, Sunnyvale, CA, USA, pp.22-27, 2016.



Thank You!

João Paulo Cabral
cabralj@adaptcentre.ie

Machine Learning Meetup
28 May 2018

