ResearchGate

See discussions, stats, and author profiles for this publication at: https://www.researchgate.net/publication/305827107

Prediction of Emotions from Text using Sentiment Analysis for Expressive Speech Synthesis

Conference Paper · September 2016

DOI: 10.21437/SSW.2016-4

CITATIONS
5

READS
361

3 authors, including:

Eva Vanmassenhove
Dublin City University
14 PUBLICATIONS **53** CITATIONS
SEE PROFILE

Fasih Haider
The University of Edinburgh
39 PUBLICATIONS **93** CITATIONS
SEE PROFILE

Some of the authors of this publication are also working on these related projects:

Metalogue View project

All content following this page was uploaded by Fasih Haider on 18 December 2017.

The user has requested enhancement of the downloaded file.

9th ISCA Speech Synthesis Workshop
13-15 Sep 2016, Sunnyvale, USA

Prediction of Emotions from Text using Sentiment Analysis for Expressive Speech Synthesis

Eva Vanmassenhove[1], *João P. Cabral*[2], *Fasih Haider*[2]

[1] Dublin City University, Ireland
[2] Trinity College Dublin, Ireland

vanmassenhove.eva@gmail.com, cabralj@tcd.ie, haiderf@tcd.ie

Abstract

The generation of expressive speech is a great challenge for text-to-speech synthesis in audiobooks. One of the most important factors is the variation in speech emotion or voice style. In this work, we developed a method to predict the emotion from a sentence so that we can convey it through the synthetic voice. It consists of combining a standard emotion-lexicon based technique with the polarity-scores (positive/negative polarity) provided by a less fine-grained sentiment analysis tool, in order to get more accurate emotion-labels. The primary goal of this emotion prediction tool was to select the type of voice (one of the emotions or neutral) given the input sentence to a state-of-the-art HMM-based Text-to-Speech (TTS) system. In addition, we also combined the emotion prediction from text with a speech clustering method to select the utterances with emotion during the process of building the emotional corpus for the speech synthesizer. Speech clustering is a popular approach to divide the speech data into subsets associated with different voice styles. The challenge here is to determine the clusters that map out the basic emotions from an audiobook corpus that contains high variety of speaking styles, in a way that minimizes the need for human annotation. The evaluation of emotion classification from text showed that, in general, our system can obtain accuracy results close to that of human annotators. Results also indicate that this technique is useful in the selection of utterances with emotion for building expressive synthetic voices.
Index Terms: expressive speech synthesis, sentiment analysis, speech clustering, emotion, audiobooks

1. Introduction

Emotional Text-To-Speech (TTS) is a challenging but important part in speech synthesis since rendering emotion makes speech sound more natural [1]. It permits to convey essential non-linguistic information that can be extracted from text in addition to the commonly modelled linguistic aspects, such as syllable stress and punctuation. This work focuses on emotional TTS for storytelling of fairy tales in audiobooks. Audiobooks have recently become a popular resource for the creation of emotional TTS systems [2]. Furthermore, fairy tales are one of the literary genres in which the emotions conveyed are of great importance for the general story line [3].

Rendering emotions in TTS is, however, not trivial. It requires solving two main problems: (a) predicting the correct emotional values of a sentence or utterance and (b) modeling and generating emotional speech [4]. Many research works focus on emotional speech synthesis [5] or sentiment analysis [6]. However, research work on the application of sentiment analysis to TTS appears to be less common. In [7], sentiment analysis is used as an input feature for expressive speech synthesis but it is only used to distinguish between different sentiment polarities (positive, negative and neutral). The expressiveness in audiobooks has a rich variety [8] and we believe that a more fine-grained distinction between emotions is necessary to better model these speech variability factors. For example, emotions belonging to the same polarity, such as 'anger' and 'sadness' (negative polarity), are characterized by different acoustic properties (intensity, pitch, speech rate, etc.), which should be modelled by the TTS system. In this work, we propose a novel emotion labelling system that uses both the information of the emotional polarity from sentiment analysis and emotion category to classify a sentence into one of the categories: anger, joy, sadness, fear, disgust, surprise and neutral.

Expressiveness in speech can be manually annotated, such as emotion annotations, or automatically classified, e.g. using unsupervised clustering techniques. For TTS, an unsupervised clustering approach is more attractive to build expressive corpora, because it is less expensive and time-consuming than creating dedicated corpora using human annotation or handcrafted rules. Recently, unsupervised clustering of expressions has been used for TTS applied to audiobooks, e.g., [9] and [10]. However, those works did not address the problem of mapping between text and clusters. Without this mapping the speech synthesis system cannot determine the appropriate expression cluster for synthesizing a given sentence. In [1], the authors propose a method for expression prediction from text and speech, in which both the expression predictor and speech synthesizer share the same training data. This method permits to model intra-speaker and inter-speaker variabilities that influence expression prediction and represent a higher number of expressions than the typical limited set of emotions of text predictor methods. In contrast to this approach, we perform the emotion prediction from text and speech clustering separately. We use the emotion labels of the sentiment analysis to automatically detect the clusters and utterances that represent each emotion. This approach has a limitation in the number of expression categories that can be detected in the audiobook compared with [1], because the latter uses a continuous expressive space which reduces the problems of labelling emotions and can synthesise speech with more detailed expressions. However, our approach has the advantage that it provides a higher degree of flexibility for manual control of the emotions of the synthetic voice and verification of emotion labels for supervised training of the classifiers. For example, the emotion labels of the utterances can be manually verified by humans to obtain more accurate speech emotion models. In this work, we developed a TTS system that uses the emotion prediction from text to select the voice style of synthesize input text.

http://dx.doi.org/10.21437/SSW.2016-4

2. Prediction of Emotion from Text

2.1. Emotional Polarity

We used the tool SentiWordsTweet [11] to distinguish sentences with a positive or negative polarity. Although this tool was developed specifically for Tweets, it has a more general application and can be used on 'regular' sentences too. The tool combines SentiWordNet [12] and AFINN [13] and introduces a new feature, called exponential weighting, which gives more importance to words appearing at the end of the sentence. SentiWordNet [12] is an opinion lexicon derived from WordNet that provides a positive, negative or objective score for every synset, while AFINN [13] is a list of 2477 words that were manually rated in terms of positivity and negativity. The sentiment-polarity score ranges between 0 (highly negative) and 1 (highly positive). In the example below, three sentences are given together with their sentiment-polarity score, sentence 1 being more negative and 3 more positive.

1. Juliet's dead | 0.354343693774
2. You're home | 0.5
3. I mean lovely | 0.586617578917

2.2. Emotional Category

In order to classify sentences according to their emotional category, we used the NRC Emotion Lexicon, which contains English words that were manually annotated by crowdsourcing [14]. Although this lexicon contains information for eight basic emotions (anger, fear, surprise, sadness, joy, disgust, anticipation and trust), we decided to only use the first six. This set of six emotional states is often used in vision and speech research and they are usually referred to as the Big Six [15]. Although these emotions might not account for all the different emotional states that can occur in audiobooks, a study by Brale et al. [16] revealed that these emotional states frequently appeared in fairy tales [17].

Since we constrained our study to the specific genre of fairy tales, we decided to extend the Emotion Lexicon to include more domain-specific vocabulary from this genre. To do so, we used a corpus of emotion annotations [17], which contains 176 fairy tales from Grimm (80 tales), H.C. Andersen (77 tales) and Potter (19 tales). We extracted the most frequent words of the fairy tales from H. C. Andersen and B. Potter and added the words related to a specific emotion that were not yet included in the lexicon. We did not use additional words from Grimm's tales because we used this data for testing in our experiments. In total, we added 208 words to the Emotion Lexicon. Table 1 indicates the number of words added for each emotion.

Anger	Disgust	Fear	Joy	Sadness	Surprise
17	21	28	90	38	14

Table 1: Number of words added to the NRC Emotion Lexicon for each particular emotion.

Our method for generating an emotion label for a sentence consists of counting the number of words in the sentence that belong to a specific emotion category. However, this counting may 'over-tag' the sentences with emotions. In order to avoid this effect, we establish a condition for our method to tag an amount of sentences with emotion similar to that obtained from human annotations. First, the highest count of emotions for a sentence is divided by the total number of tokens in the sentence. If this division results in a number higher than a threshold, the sentence is labelled with the top rated emotion for that sentence. The threshold of 0.07 was obtained by matching the amount of automatically 'emotion-tagged' sentences using the sentiment analysis tool and the human annotations of emotion from [17].

We also added an extra condition in the automatic emotion labelling by only allowing the sentence to be labeled into a specific category if the sentiment-polarity (positive or negative) corresponds to the polarity of the emotion. For this decision, the sentiment polarity was obtained from the sentiment analysis score (range of 0 to 1, with positive polarity being higher than 0.5 and negative lower than this threshold), while the emotions were divided into negative (anger, disgust, fear, sadness, surprise) and positive (joy). The examples of sentences below help to explain the emotion labelling based on sentiment polarity. In sentences (1) and (2) the emotion-label agrees with the sentiment-polarity (negative polarity for 'fear' and positive polarity for ''joy') while in (4) and (5) this is not the case (negative polarity for joy and positive polarity for fear). Thus, the emotion labels of (4) and (5) become 'neutral'.

1. "Juliet's dead" | 0.35 | fear
2. "I mean lovely" | 0.59 | joy
3. "You're home" | 0.5 | neutral
4. "What name did they give the child?" | 0.44 | joy - NEUTRAL
5. "May God help you with your fishing" | 0.65 | fear - NEUTRAL

3. Emotional Speech Corpus

3.1. Audiobook Corpus

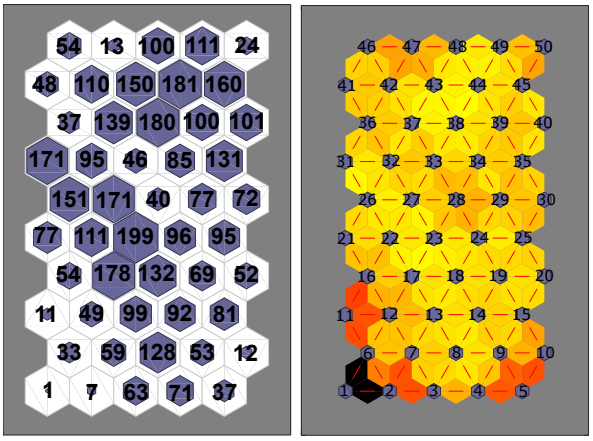
We used the dataset of audiobooks released for the Speech Synthesis Blizzard Challenge 2016 (`http://synsig.org/index.php/Blizzard_Challenge_2016`). It consists of speech and text data of professional audiobooks and includes about 5 hours of British English speech data (sampled at 44 kHz) from a single female talker. However, for part of this dataset the tales were provided in pdf format. We excluded these audiobooks because we found problems in the extraction of the text from the pdf files and to avoid the effect of any errors in the text extraction. Thus, we only used the subset of the database that consists of 25 fairy-tales (half the total number of the audiobooks of the Blizzard dataset).

Although sentence-level alignments between speech and text are included in the dataset for part of the data, we used our own alignments that were obtained with the Kaldi toolkit (`http://kaldi-asr.org`). The new alignments allowed us to segment the data into direct speech and narrator speech. The sentences corresponding to direct speech were obtained by automatically detecting the quoted text. We observed that in the data, direct speech was generally more expressive (and thus more 'emotional') than the narrator parts (tend to be more 'neutral'). For this reason, we selected the emotional utterances from the direct speech subset. We considered two different ways of selecting the group of utterances that represent each emotion. The first employs speech clustering and the other is based solely on the sentiment analysis where utterances with a 'stronger' sentiment analysis score are selected (Section 3.3).

3.2. Clustering of Speech Styles

We used Self Organising Map (SOM) to cluster the speech data of the audiobooks, similarly to [10], which was implemented with the MATLAB Neural Networks toolbox. The acoustic feature set was extracted using the openSMILE toolkit [18]. We initially used the same feature set that was used in openEAR [19] to recognize emotions in real time. For example, it includes spectral parameters (MFCC, LSP, etc.), intensity and pitch related parameters. The configure file (emobase_live4.conf) for feature extraction can be downloaded from `https://sourceforge.net/projects/openart/`. We did a correlation test between these features and the duration of the speech utterances in order to remove features which depend on duration (we removed those features whose correlation was greater than 0.2). Consequently, we reduced the size of the feature set from 950 to 605.

We performed the clustering of the data with different numbers of clusters ranging from 25 to 50. The reason for varying the number of clusters was to verify that this number had an impact on the variability of voice styles observed between the clusters. We assumed that by using a sufficiently high number of clusters, we could obtain clusters that represent a particular emotion. However, with the increase in the number of clusters, there is higher possibility that different clusters may be characterized by similar voice styles. From informal listening tests performed by the authors we decided that 50 clusters was the best option for obtaining a good separation of speech emotions. Figure 1 shows the SOMs obtained for this condition.



(a) Cluster size = 50 (b) Cluster size = 50

Figure 1: The left plot shows the number of utterances in each cluster and the right plot shows the distance between the numbered clusters (the higher the distance the darker the colour).

In order to map speech clusters to emotions, we performed sentiment analysis on the sentences belonging to a specific cluster and compared its emotional distribution with the overall distribution of emotions (over all the clusters). For example, Figure 2 shows that the cluster number 11 has a significantly high rate of labels 'anger'. Thus, this cluster is expected to have a high amount of utterances that sound with this emotion. We conducted preliminary experiments to test this assumption and the results are positive as it will be shown later. By using the information about the distance between clusters (Figure 1) together with the scores of the sentiment analysis it is possible to automatically select a set of candidate clusters to build the synthetic voices for each emotion. For example, the clusters 11, 4, 12, 16, and 28 have relatively high distances to the other clusters and obtained high classification rates for the emotions 'anger', 'sadness', 'joy', 'fear', and 'surprise', respectively.

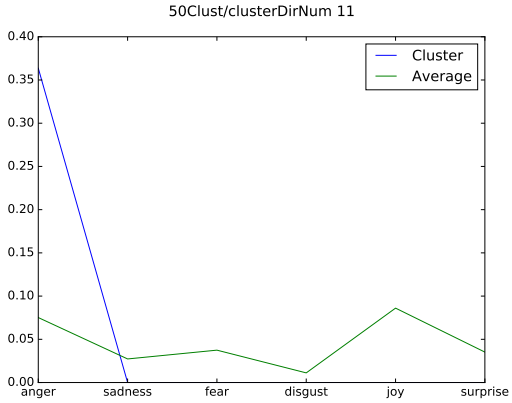


Figure 2: Distribution of emotions for the cluster 11 compared to the average distribution of emotions over all clusters.

3.3. Selection of Utterances with Emotions using Sentiment Analysis

We also predicted the emotions of the utterances based solely on the sentiment analysis of the text, which consists of making the condition for emotion classification more restrictive. As described in Section 2.2, we only allowed sentences to be labeled with a specific emotion if the sentiment-polarity score corresponded to the polarity of the emotion. Thus, we can control the threshold applied to the sentiment-polarity values in order to control the number of sentences labelled with emotion and to select sentences with stronger 'sentiment level' for that emotion. For example, we can select only sentences for which the sentiment-polarity score is lower than the threshold of 0.35, which results in 10 strongly negative 'sad' sentences.

4. Expressive TTS system

4.1. HMM-based Speech Synthesizer

We used the HTS system [20] (version 2.3), available at `http://hts.sp.nitech.ac.jp`, to build our synthetic voices. This is a popular system which permits to build voices for a target voice style using a relatively small amount of speech data, by using adaptation techniques [21].

For speech analysis, the STRAIGHT method [22] (Matlab version V40_006b) was used to calculate the spectral envelope of the speech signal and the aperiodicity measurements, while F_0 was calculated using the RAPT algorithm implementation of the ESPS tools [23], [24]. The Fast Fourier Transform (FFT) parameters of the envelope and aperiodicity are converted to 39th order mel-cepstral coefficients, while the FFT parameters of aperiodicity are weighted in 25 frequency bands. HTS performs this parameter conversion to obtain a parameter representation which is more compact and better for statistical modeling.

Acoustic modeling was performed using the standard five-state left-to-right MSD-HSMM structure and both the state output density function and the state duration were modelled by

a single Gaussian distribution. The feature vector consists of five streams: mel-spectrum, aperiodicity parameters, F_0 and its Δ and Δ^2 parameters. The spectrum and aperiodic feature vectors consist of their static and dynamic parameters (Δ and Δ^2), respectively. HTS also performs clustering of context-dependent HMMs using decision trees. We used the text analysis component of the Festival system (`http://festvox.org/festival/`) to extract the linguistic features and generate the context labels.

During the synthesis stage, the input text is analyzed to generate linguistic labels using Festival and the emotion of each sentence is obtained with the emotion prediction component described in Section 2. The emotion is used to select one of the voices with emotion which were built using an adaptation technique. STRAIGHT is also used here to generate the speech parameters obtained from the HMMs and linguistic labels.

4.2. Synthetic Voices

First, an average voice model was built with the HTS system using all the speech and text data of the audiobook corpus described in Section 3.1. Then, the resulting HMM models were adapted to each target emotion voice using the MLRR adaptation method and the corresponding subset of speech with emotion that was obtained from the corpus using the following semi-automatic technique. First, we obtained around 50 utterances for each emotion using the technique explained in Section 3.3. For some emotions, e.g. disgust, the total number of labelled sentences was lower than 50 (we evaluated as many as provided in the data in these cases). Next, one of the authors listened to those samples to verify if the emotion predicted from the text was in concordance with the emotion perceived from speech. From this analysis, we selected at least 20 utterances to represent each speech emotion (ranged from 26 to 54 utterances depending on the amount of correct labels available for each emotion). In this work, we did not use speech clustering for selecting the emotion datasets (as described in Section 3.2), because for some emotions the number of utterances obtained from mapping the clusters to emotions was small (lower than 20 utterances). We plan to do experiments using this technique by using a larger speech corpus. Finally, after performing the adaptation using HTS, we obtained seven synthetic voices for the emotions: joy, anger, sadness, surprise, disgust, fear and neutral. Examples of speech synthesized with these emotions are available online (`https://www.scss.tcd.ie/~cabralj/samples-emotion-tts.html`).

Other possible approaches could be used to integrate the emotion prediction into the HMM-based speech synthesizer. One technique would be to augment the linguistic labels to include the emotion information and to train only one synthetic voice. This voice should be able to produce the desired variations in speech emotion depending on the input text. However, the rate of non-neutral emotion labels is significantly low as indicated later in Table 2 and we think it would be difficult to model well those variations using one single voice model. Moreover, the high number of linguistic labels would also contribute to the problem of modeling the expressiveness aspect of the voice on the clustered HMM models. Another technique would be to try to predict the emotions from text using HTS by embedding the emotion prediction into the decision tree stage. We also did not considered this option here because we assume that an emotion classifier based on the decision tree technique requires a much larger amount of emotion labelled data than that available in this work. These two alternative options that only require training of one synthetic voice have the advantage that they can be easily scaled to the synthesis of additional voice types without the need for building and selection different voices. Their great disadvantage compared with a separate emotion prediction component is the impact of linguistic factors on the expression prediction performance.

Emotion labels	Annotators	Proposed System
Anger	4.1%	2.3%
Sadness	3.4%	2.8%
Joy	6.2%	7.4%
Fear	2.9%	1.9%
Surprise	1.6%	3.2%
Disgust	0.3%	0.8%
Total for emotions	18.5%	18.4%
Total for neutral	81.5%	81.6%

Table 2: Relative amount of emotion-labels per emotion in the human annotations and obtained by our system.

5. Results

5.1. Prediction of Emotion from Text

For evaluating our emotion classification tool we used the human annotations of the Grimm fairy tales from [17]. We intentionally excluded the tales of Grimm in the extension of the emotion dictionary in order to be able to test how our system works with unseen data. In the annotated data, every sentence is labelled with a primary emotion and a type of mood provided by two annotators, resulting in a total of 4 affect labels per sentence [17]. Table 2 shows the relative amount of labels for the emotional categories obtained from the human annotations and from our emotion-labelling method. We only counted the sentences as belonging to a particular emotion if both annotators agreed on the primary emotion. From the table, we observe that most of the sentences in the fairy tales are neutral (82%) and 'joy' is the most frequent emotion both in the human annotations and in the predicted emotion labels. We also observe that the total number of sentences labelled with emotion is approximately the same between the two methods. This result shows that our system can be effectively used to approximate the amount of emotion labels predicted from text to that of human annotations.

The F-score of the emotion prediction tool is calculated based on the precision and recall and the results are shown in Table 3. We consider a sentence as correctly tagged by the tool if the detected emotion label matches any of the four affect labels of the human annotations. For each emotion, the precision is defined by the number of correct labels for the emotion divided by the total number of labeled sentences, while recall is the number of correct labels for the emotion divided by the number of labels of that emotion obtained in the human annotations. The precision of the tool over all sentences was 79%. The F-score was highest for the neutral sentences (0.86), while for sentences labeled with a particular emotion, 'joy' had the highest F-score (0.38).

In Table 3, we also present the F-score for one human annotator. The score is based on the sentences in which both annotators agreed on, compared to the total of sentences tagged. The averaged precision and recall of both annotators were 55% and 50% respectively, resulting in an average F-score of 52%.

Emotions	F-Score System	F-Score Annotator 1
Anger	0.20	0.41
Sadness	0.30	0.39
Joy	0.38	0.53
Fear	0.18	0.32
Surprise	0.09	0.38
Disgust	0.00	0.09
Neutral	0.86	0.71
Average	**0.30**	**0.48**

Table 3: Comparison of F-scores between emotion-labelling tool and one annotator.

Emotion Labels	Prediction for text	Prediction for speech	Prediction for speech (text labels correct)
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%**

Table 4: Results obtained from the human evaluation of the emotion labels for text and speech.

Although the F-scores of the emotion prediction tool are not as high as for human annotations, the analysis of the human labels gives support to the assumption that fine-grained sentiment analysis is a very difficult and sometimes subjective task. We did not find in the literature directly comparable results for sentence-level prediction of the same 6 emotions in terms of F-score, precision and recall. In future work, we plan to compare our approach of emotion prediction from text with other methods that are not lexicon-based by using the same dataset.

5.2. Classification of Speech Emotions from Text

We evaluated the emotions of the data subsets that were automatically obtained using sentiment analysis for building the synthetic voices (Section 4.2). In this experiment, one of the authors listened to approximately 50 utterances labelled by the tool for each emotion. The results of our human-based evaluation are presented in Table 4. The second column of the table indicates the rate of labels of the tool that are correct according with the human evaluation, by considering the text only without any context. Meanwhile, the third column shows how often the emotion predicted from text was the same as that perceived by the listener. Finally, from these results we also calculated the rate of emotions perceived by the listener that matched the emotion predicted from text by considering the correct labels only (from the text evaluation). These results indicate that there is strong correlation between the emotion predicted from text and the corresponding speech emotion, especially for the emotions 'surprise' and 'anger'. These preliminary results give support to the hypothesis that our emotion prediction system can be useful in automatic selection of speech with emotions for building expressive synthetic voices. However, more extensive experiments with more data and a larger number of listeners are needed to obtain more conclusive results about this hypothesis.

6. Conclusions and Future Work

We developed a method that combines the information of fine-grained lexicon-based sentiment analysis with sentiment-polarity scores in order to get more accurate emotion labels. The resulting emotion prediction tool was integrated into an HMM-based speech synthesizer to select one of the following types of synthetic voice from the input text: anger, fear, surprise, joy, sadness, disgust and neutral. We showed that the predictions of emotion from text using our method were generally close to those obtained by human annotation, with exception of some emotions which also obtained lower agreement between annotators (particularly for disgust, surprise and fear). Our emotion prediction tool also permits to have control over the number of sentences labelled with emotion by using a threshold of sentiment polarity score. This allowed us to obtain a number of emotional labels similar to human annotation in the audiobook dataset that we used for speech synthesis. We also used this threshold-based technique to select small subsets of the audiobook data with strong emotion scores for building the synthetic voices. However, results of a preliminary perceptual experiment showed that the textual sentiment analysis does not always correspond to the emotions conveyed in uttered speech. Nevertheless, we assumed that this correspondence was high enough to obtain convincing expressive voices. We are conducting a perceptual experiment to further investigate this question and evaluate if the TTS system developed in this work conveys speech emotions correctly. We also plan to compare our sentiment analysis method to different approaches, in particular non-dictionary based methods.

7. Acknowledgements

This research is supported by the Science Foundation Ireland (Grant 13/RC/2106) as part of ADAPT (www.adaptcentre.ie) and EU FP7-METALOGUE project under Grant No. 611073, at Trinity College Dublin, and by the Dublin City University Faculty of Engineering & Computing under the Daniel O'Hare Research Scholarship scheme.

8. References

[1] Chen, L., Gales, M., Braunschweiler, N., Akamine, M. and Knill, K., "Integrated Expression Prediction and Speech Synthesis From Text", IEEE Journal of Selected Topics in Signal Proc., 8(2):323–335, 2014.

[2] King, Simon, and Vasilis Karaiskos. "The blizzard challenge 2012." (2012).

[3] Alm, Cecilia Ovesdotter, Dan Roth, and Richard Sproat. "Emotions from text: machine learning for text-based emotion prediction." Proceedings of the conference on human language technology and empirical methods in natural language processing. ACL. 2005.

[4] Cahn, Janet E. "The generation of a ect in synthesized speech." Journal of the American Voice I/O Society 8 (1990): 1-19.

[5] Schröder, Marc. "Emotional speech synthesis: a review." INTERSPEECH. 2001.

[6] Agarwal, Apoorv, Fadi Biadsy, and Kathleen R. Mckeown. "Contextual phrase-level polarity analysis using lexical affect scoring and syntactic n-grams." Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics. ACL, 2009.

[7] Trilla, Alexandre, and Francesc Alias. "Sentence-based sentiment analysis for expressive text-to-speech." Audio, Speech, and Language Processing, IEEE Transactions on 21.2 2013. pp. 223–233.

[8] Charfuelan, Marcela, and Ingmar Steiner. "Expressive speech synthesis in MARY TTS using audiobook data and emotionML." INTERSPEECH. 2013.

[9] Eyben, F, Latorre, J., Wan, V., Gales, M. and Knill, K., "Unsupervised clustering of emotion and voice styles for expressive TTS", ICASSP, pp. 4009 - 4012, 2012.

[10] Székely, E., Cabral, J. P., Cahill, P. and Carson-Berndsen, J., "Clustering expressive speech styles in audiobooks using glottal source parameters", Proc. of Interspeech, Florence, Italy, 2011.

[11] Afli, Haithem et al. "SentiWordsTweet". adapt-invention disclosure form, 2016 (in Press).

[12] Miller, George A. "WordNet: a lexical database for English." Communications of the ACM 38.11 1995. pp. 39–41.

[13] Hansen, Lars Kai, et al. "Good friends, bad news-affect and virality in twitter." Future information technology. Springer Berlin Heidelberg, 2011. pp. 34–43.

[14] Mohammad, Saif M., and Peter D. Turney. NRC Emotion Lexicon. NRC Technical Report, 2013.

[15] Cornelius, Randolph R. "Theoretical approaches to emotion." ISCA Tutorial and Research Workshop (ITRW) on Speech and Emotion. 2000.

[16] Brale, Vronique, et al. "Towards an expressive typology in storytelling: A perceptive approach." Affective Computing and Intelligent Interaction. Springer Berlin Heidelberg, 2005. pp. 858–865.

[17] Alm, Ebba Cecilia Ovesdotter. Affect in text and speech. ProQuest, 2008.

[18] Eyben, F., Wöllmer, M. and Schuller, Björn. "Opensmile: the Munich versatile and fast open-source audio feature extractor", Proc. of the 18th ACM international conference on Multimedia, pp. 1459–1462, 2010.

[19] Eyben, F., Wöllmer, M. and Schuller, Björn. "OpenEAR introducing the Munich open-source emotion and affect recognition toolkit", Affective Computing and Intelligent Interaction and Workshops, pp. 1–6, 2009.

[20] Tokuda, K., Masuko, T., Miyazaki, N. and Kobayashi, T., "Hidden Markov models based on multi-space probability distribution for pitch pattern modeling", Proc. ICASSP, pp. 229–231, 1999.

[21] Tachibana, M., Yamagishi, J., Onishi, K., Masuko, T., Kabayashi, T., "HMM-based speech synthesis with various speaking styles using model interpolation", Proc. of Speech Prosody, 2004.

[22] Kawahara, H., Masuda-Katsuse, I. and Cheveigné, A., "Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based f_0 extraction: Possible role of a repetitive structure in sounds", Speech Communication, Vol. 27, pp. 187–207, 1999.

[23] Talkin, D., "Voicing epoch determination with dynamic programming", J. Acoust. Soc. Amer., 85, Supplement 1, 1989.

[24] Talkin, D. and Rowley, J., "Pitch-Synchronous analysis and synthesis for TTS systems", Proc. of the ESCA Workshop on Speech Synthesis, C. Benoit, Ed., Imprimerie des Ecureuils, Gieres, France, 1990.

View publication stats