IberSPEECH 2018
21-23 November 2018, Barcelona, Spain

# Unsupervised Learning for Expressive Speech Synthesis

*Igor Jauk*[1]

[1]Universitat Politècnica de Catalunya, Spain
ij.artium@gmail.com

## Abstract

This article describes the homonymous PhD thesis realized at the Universitat Politècnica de Catalunya. The main topic and the goal of the thesis was to research unsupervised manners of training expressive voices for tasks such as audiobook reading. The experiments were conducted on acoustic and semantic domains. In the acoustic domain, the goal was to find a feature set which is suitable to represent expressiveness in speech. The basis for such a set were the i-vectors. The proposed feature set outperformed state-of-the-art sets extracted with OpenSmile.

Involving the semantic domain, the goal was first to predict acoustic features from semantic embeddings of text for expressive speech and to use the predict vectors as acoustic cluster centroids to adapt voices. The result was a system which automatically reads paragraphs with expressive voice and a second system which can be considered as an expressive search engine and leveraged to train voices with specific expressions.

The third experiment evolved to neural network based speech synthesis and the usage of sentiment embeddings. The embeddings were used as an additional input to the synthesis system. The system was evaluated in a preference test showing the success of the approach.

**Index Terms**: expressive speech, unsupervised training, semantic embeddings, acoustic features

## 1. Introduction

Speech synthesis is an old, almost romantic, idea of machines, computers, and robots, who talk and express themselves as do human beings. In futuristic science fiction movies and literature, there is almost no way around a talking computer or robot. Sometimes, the talking computer, despite of the vast artificial intelligence capabilities, is identified as such, talking in a robotic and monotonous way, and being nothing else than an aid to humans. In different occasions, computers act very human-like, imitating emotions and free will. In today's world of smartphones, mobile connectivity and fast-paced life, speech applications gain more and more importance – not least due to ever-improving synthetic speech quality. A fast look at these applications reveals clearly that today's users do not looking for an AI with a robotic and monotonous voice: making jokes, understanding and showing sympathy, and a long etc. of things which can be resumed in "sounding human-like", almost seems to be a warrant for product quality of such applications.

In a not too far past, *expressive speech* (or also *emotional* speech) was pretty much of an effort. Systems designed specific emotions, often with control mechanisms for intensity etc., and achieved impressive results with such techniques – e.g. [6, 28, 5, 13, 41], just to name a few. However, despite the partly great sounding results, and apart of the laborious design, these systems focus on a limited set of emotions or expressive speaking styles. Real-life human speech does include an infinite number of emotions as those are speaker and situation dependent. So in order to create a truly flexible system you need a mechanism which can adopt to all possible conversational situations and create expressions "on the fly". For this we need automatic processes for analysis of data and training. Some suggestions to approach this problem were made for instance in [11, 39, 7, 23], where data was clustered by different criteria to select training data and similar ideas. And this is also the topic of this work. The guiding hypothesis are:

1. It is possible to define expressive voices from clusters of data in the acoustic domain, applying unsupervised methods to build the clusters, i.e. no labels of human interpretation are permitted to define the voices or the data in the clusters.
2. It is possible to improve the expressiveness of a synthetic voice using in the training process semantic features which codify some sort of expressive information and are obtained fully automatically.

So one of the main goals is the authomatic processing. Also, as the hypothesis reveal, the work embraces both, the acoustic, and the semantic domains. The different approaches are presented in the sections below. Section 2 introduces *i-vector-based* representation methods for expressive speech. Section 3 introduces prediction methods of acoustic features from semantic text representations, also called *embeddings*. Section 4 discusses a neural network based TTS system which uses *sentimente embeddings* to "predict" expressiveness. Finally, Section 5 provides a small discussion and draws some conclusions, and Section 6 lists the helping parties of this project.

## 2. Acoustic feature selection

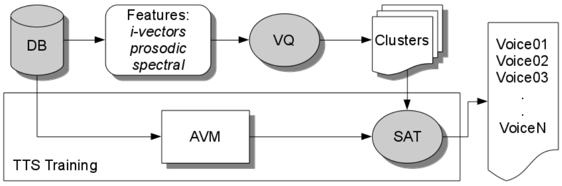
Traditionally, expressive speech synthesis and analysis leveraged especially prosodic features to represent emotions or speaking styles. For instance [19] use F0, intensity and duration; [35] use glottal parameters. In [32] the authors use a set of prosodic parameters combined with some spectral ones; [11] use prosodic features, i.e. F0, voicing probability, local jitter and shimmer, and *logarithmic HNR* for audiobook clustering and posterior synthetic voice training. Almost all research concentrate on prosodic features. However, there are also findings which state that some emotions might be better represented with spectral parameters, e.g. [27, 1]. In [22] the authors used i-vectors (e.g. [31, 8]) to predict emotions. This idea was picked up in [17, 16] and developed to a set of features, i-vector-based and others, which were compared in three clustering experiments.

### 2.1. Experimental framework

The first and the second experiments compared multiple features objectively and subjectively, as published in [17, 16]. The framework is presented in figure 1.

10.21437/IberSPEECH.2018-38

Figure 1: *Clustering and synthesis framework.*



The idea is: use different feature sets to cluster expressive speech, use the data in clusters to train expressive voices and synthesize a diologue using these expressive voices. The corpus is a juvenile narrative audiobook recorded in European Spanish, with a total of 7900 sentences and 8.8 hours of duration. The clustering algorithm is k-means, concretely *VQ*, as by [12]. Many different featur combinations were tested. The results of some of them a presented below. Some features were combined to sets in order to facilitate the notation.

- *silRate* is silence rate and *syllRate* is syllable rate (#/sec) (extracted with *Ogmios* [4]).
- *Rhythm* is silence and syllable rates (#/sec), duration means and variation, computation based on segmentation.
- *Pitch* is F0 means, variance and range.
- *JShimm* is Local jitter and shimmer (extracted with [3]).
- *MFCCiVec* are i-vectors calculated on basis of MFCCs; *F0iVec* are i-vectors calculated on basis of F0. *iVecC*: F0 and MFCC based i-vectors (the acoustic features for the i-vectors were extracted with the AHOCoder [9], the i-vectors themselves were extracted using *Kaldi* [29]).

## 2.2. Objective evaluation

An objective evaluation was performed: a small part of the corpus was labeled with expressions and character (speaker) labes (only for the evaluation purposes) and with the aid of these labels, the perplexity of the clusters was calculated, derived from entropy as by [33, 42]: $PP = 2^{\bar{H}(X)}$.

A resume of the results is shown in the table 1. Due to space limitations only few chosen results are shown. The upper line shows the perplexity calculated for the annotated part of the corpus. The part below it shows the performance of some "traditional features" and the part at the bottom, the performance of sets which included i-vectors.

As can be seen, the combination of Rhythm and iVecC outperformed all other combinations in the given task.

Table 1: *Perplexities (PP) for different feature sers for Expressions (Ex) and Characters (Ch) in comparison to the database.*

| | PP/Ex | PP/Ch |
|---|---|---|
| $DB$ | 140.4 | 8.3 |
| $silRate$ | 10.6 | 4.7 |
| $sylRate$ | 9.4 | 4.0 |
| $meanF0$ | 9.8 | 4.2 |
| $Rhythm - Pitch$ | 8.7 | 3.4 |
| $Rhythm - Pitch - JShimm$ | 8.6 | 3.4 |
| $MFCCiVec$ | 9.0 | 3.5 |
| $F0iVec$ | 11.2 | 4.2 |
| $iVecC$ | 8.8 | 3.8 |
| $Rhythm - iVecC$ | **8.2** | **3.3** |
| $Rhythm - JShimm - iVecC$ | 8.5 | 3.5 |

In order to verify these results, an additional objective evualation was conducted. For this evaluation, *OpenSmile* feature sets, as in *openSMILE Book* [10], were compared to the proposed sets. OpenSmile is a set of feature extraction tools widely used for emotional and expressive speech analysis and synthesis. It extracts thousands of features and statistics about them and is considered to be state-of-the-art feature extraction for expressive speech. The table 2 compares some OpenSmile feature sets to the winning i-vector set: Rhythm & iVecC.[1] Also here, the proposed combination of Rhythm & iVecC outperformed all OpenSmile sets.

Table 2: *Perplexities for different features combinations, including openSMILE for expressions (E) and for characters (Ch).*

| | PP/Ex | PP/Ch |
|---|---|---|
| $DB$ | 140.4 | 8.3 |
| $is09$ | 10.5 | 3.9 |
| $is10$ | 10.8 | 4.0 |
| $emobase$ | 10.5 | 4.0 |
| $emolarge$ | 8.8 | 3.7 |
| $Rhythm - iVecC$ | **8.2** | **3.3** |

## 2.3. Subjective evaluation

Two subjective experiments were conducted. For these experiments, for a given dialogue from the same audiobook (test set excluded form the training set), a set of synthetic voices was trained using the data in the clusters. The underlying system was an HMM based TTS [24] where the average voice was trained using the whole corpus (aprox. 10h) and the cluster data was used to perform adaptation. A total of 16 sentences was presented to a number of participants. The task: design your own audiobook dialogue using synthetic voices instead of the real once.

The interface is a website, where the participants could choose 1 of 10 synthetic voices for each sentence in a diolague. The website design aimed to create the right atmosphere of the book story and a more enjoyable experience. Also an introduction text was provided for the case that the participants were not familiar with the story.

The experiments differ in so far, that in the first experiments the participants had an example of the original character voice, in the second they did not. Also: in the first experiment, the synthetic voices were chosen manually with the criterion of resamblance to the original voices, and mixed with other random voices. In the second that choices was made automatically by acoustic distance for half of the sentences, the rest was random.

[1] *is09, is10, emobase, emolarge* are feature sets by OpenSmile used in different experiments. For further details please refer to *openSMILE Book* [10].

The idea behind: if some voices are especially suitable for some characters, the participants would tend to prefer them.

In the first experiment, 19 persons had participated; in the second, 11 persons. Due to space limitations only the results for the second experiment are shown in Table 3.

Table 3: *Relative preferences for the voices v0-v9 over the whole paragraph for the narrator (Narr) and the two present characters (Ch2 and Ch3).*

| | v0 | v1 | v2 | v3 | v4 | v5 | v6 | v7 | v8 | v9 |
|---|---|---|---|---|---|---|---|---|---|---|
| $Narr$ | **0.42** | 0.06 | 0.00 | 0.03 | 0.04 | **0.23** | 0.04 | 0.10 | 0.06 | 0.01 |
| $Ch2$ | 0.13 | **0.16** | 0.14 | **0.23** | 0.03 | 0.09 | 0.05 | 0.03 | 0.10 | 0.03 |
| $Ch3$ | **0.18** | 0.13 | 0.13 | **0.31** | 0.00 | **0.18** | 0.00 | 0.00 | 0.00 | 0.05 |

The results show that there is an actual preference for some voices atop of others. Of course, not all participants have the same imagination of the book characters, especially if they don't know the book. So individual preferences are out of scope of this task. But the results show clearly, that the approach as well as the proposed feature sets are suitable for the task. The task itself can be interpreted as a simulatio of a real-life application of expressive speech. Further details on the experiments and the results are published in [17, 16, 14]

## 3. Semantics-to-Acoustics Mapping

When we humans read a text aloud –lets say we read a good night story to a small child–, we probably will read the story with an expressive voice imitating book characters, their emotions in different situations etc. as to engage the child. Although likely we all would read slightly different, though the "quality" of our reading will possibly be judged by our expressive abilities. The question for this section is: What in the text does provide us the necessary information as to adequately adapt our reading style and can it be taught to a machine?

The key approach is the automatic representation of text. Such representations are often called *embeddings* and there is a large number of techniques to calculate them, like for instance [21, 2, 26, 34]. The basic idea is to represent text in terms of word co-occurences of defined text units (e.g. sentences, paragraphs, etc.). This way each unit is represented as a co-occurence vector of its own words. More modern techniques train neural networks to predict a certain feature. Then extract the vector representations from intermediate layers of the network, like [26] and [34].

The assumption is that these representations actually also codify expressive information, especially if the underying network is trained using an "expressive" criterion (see Section 4). So the task is to convert this vector representation into acoustics.

### 3.1. Experimental framework

The framework is: in a given text corpus, in this case an audiobook, for each sentence of the text a semantic embedding is calculated. This embedding is then used to predict an acoustic feature vector, concretely from the above experiment, which for its part is the centroid of a data cluster. As in the experiments above, these data clusters are used for adaptation in an HMM TTS. The performance of the system is evaluated in two subjective experiments. For the embeddings the toolkit *word2vec* [40, 25] was used to calculate the word embeddings; the sentence embeddings were calculated as centroids of the word embeddings in the vector space. The vector space has been trained with the *Wikicorpus* [30].

### 3.2. Subjective evaluation

The task in the first experiment was to read two book paragraphs automatically predicting expressiveness for each sentences. For each sentence in the paragraph an embedding was calculated. It was used to predict an acoustic feature vector as in Section 2. Two prediction models were compared: a nearest-neighbour classifier and a neural network. Both paragraphs were extracted from different books of the same series, as to preserve characters and the ambience. The expressive readings were also compared to a neutral reading. The task was implemented as a preference test. The participants had the option to choose that two systems performed equally.

A total of 21 persons participated in the experiment. Table 4 shows the results for these experiment for both paragraphs ($P1$ and $P2$). The results show a clear preference for the expressive systems.

Table 4: *Prediction method preferences by users for the first two tasks. DNN method, nearest neighbor (NN) method, neutral voice.*

| | | | | DNN | NN |
|---|---|---|---|---|---|
| | DNN | NN | neutral | =NN | =neutral |
| $P1$ | 0.19 | 0.43 | 0.0 | 0.38 | 0.0 |
| $P2$ | 0.29 | 0.14 | 0.04 | 0.48 | 0.05 |

In a further test the prediction system was used as a "search engine" for expressive training data. For this purpose, a keyword called *seed* was used to predict an acoustic vector, which on its side was used as a cluster centroid for acoustic data and for voice adaptation. For example "Mysterious secret in silent obscurity" was used as seed to find training data for a *suspense* voice. Other trained emotions were *angry, happy* and *sadness*, also a neutral voice. Seven sentences were synthesized with each of these voices and again, a preference test was presented to the 21 participants. The synthesized sentences were chosen trying to reflect their expressive meaning. For example "Finally, the holidays begin!" is supposed to be happy and the expectation was that the participants would choose the happy voice for it. Table 5 presents the results for this experiment.

Table 5: *Task 3. Voice preference by users for each sentence.*

| | happy | angry | suspense | neutral |
|---|---|---|---|---|
| Happy$_1$ | **0.29** | **0.38** | **0.24** | 0.10 |
| Happy$_2$ | **0.52** | 0.24 | 0.10 | 0.14 |
| Angry$_1$ | 0.14 | **0.48** | 0.24 | 0.14 |
| Angry$_2$ | **0.38** | **0.43** | 0.14 | 0.05 |
| Suspense | 0.0 | 0.05 | **0.81** | 0.14 |
| Sadness | 0.19 | 0.05 | **0.43** | **0.43** |
| Neutral | 0.10 | 0.05 | **0.43** | **0.43** |

The preferences for voices are pretty clear. It is interesting to remark that happy and angry voices for sometimes exchangeable, the same can be said about the sad and neutral voices in their respective contexts. Further details on these experiments can be found in [15, 14].

## 4. NN-based expressive TTS with sentiment

Looking back at the experiment in the previous chapter there are few spontaneous suggestions which can be made as to develop the approach and improve the results. First, nowadays HMM-based synthesis is almost completely replaced by synthesis based on neural networks. NN-based TTS provides new possibilities of leveraging semantic vectors and avoiding clustering, which is an advantage itself since all data is always taken into account in the training process. For this experiment the DNN-based TTS as described in [36] was used.

The second point is the usage of embeddings which are more suitable to represent expressiveness. The authors in [20, 34] propose the Stanford Sentiment Parser. The system is trained on movie reviews and predicts the positivity or negativity of the sentence.

In previous work, neural network based systems have already been combined with semantic vector input, though not for expressive speech. To name a few, [37] use word embeddings to substitute TOBI and POS tags in RNN-based synthesis achieving significant system improvement. [38] enhance the input to NN-based systems with continuous word embeddings, and also try to substitute the conventional linguistic input by the word embeddings. They do not achieve performance improvement, however, when they use phrase embeddings combined with phonetic context, they do achieve significant improvement in a DNN-based system. [38] enhances word vectors with prosodic information, i.e. updates them, achieving significant improvements.

### 4.1. Experimental framework

The DNN-based system was trained on two audiobooks in American English. An additional linguistic input is introduced, the sentiment predicted by the Stanford sentiment parser. Here, different input combinations are tested, including word context.

With word context, probability vector of the word in question and the probability vectors of two words on the left and two words on the right were used. Also the tree distance, which is the hierarchical distance counted in the number of binary tree nodes between words is added, such that the input vector for each word for the system (v_wcd) is composed as follows:

$$P = \{P_{l_2}, P_{l_1}, P_c, P_{r_1}, P_{r_2}, D_t\} \quad (1)$$

where $P_c$ is the probability vector for the current word, the $P_{l_2}$ is the probability vector for the second word on the left, $P_{l_1}$ is probability vector for the first word on the left, $P_{r_1}$ is probability vector for the first word on the right and $P_{r_2}$ is the probability vector for the second word on the right, each of the probability vectors as defined in equation **??**. $D$ is the hierarchical tree distance (distance in tree counted in nodes).

### 4.2. Subjective experiments

Two experiments have been conducted in order to test the system performance. In all of them, similar to the experiment above, a preference test is conducted among a group of participants. A total of 20 persons participated in the experiment. However, there was a larger group of speech technology experts and people who have no experience with speech technology which allowed for an interesting comparison of the deviations of these two groups. Due to space limitations only the general preference results will be shown here. More details can be found in [14, 18].

Table 6 shows the preferences divided by the sentiment. For positive and negative sentences, the *word level* system performed best, although for negative sentences with high variance. For neutral sentences, the *word context and tree distance* system performed best. Possibly it is due to the fact that it probably has an equilibrating effect.

T-tests show that for negative sentences, there is a significant difference between the system without sentiment and the *word level* system, and no significant difference for the other systems. For neutral sentences, there is a significant difference between the system without sentiment and the *word context and tree distance* system, but not for the other systems. For positive sentences, there is only significant difference for the one-tailed t-test between the system without sentiment and the *word level* system.

Table 6: *System preferences for positive, negative and neutral sentences. ws: without sentiment, wcd: word context and tree distance, wl: word level*

| | ws | wcd | wl |
|---|---|---|---|
| positive mean | 1.84 | 1.85 | 1.71 |
| positive variance | 0.54 | 0.76 | 0.54 |
| negative mean | 2.06 | 1.96 | 1.84 |
| negative variance | 0.52 | 0.67 | 1.1 |
| neutral mean | 2 | 1.83 | 1.96 |
| neutral variance | 0.71 | 0.6 | 0.95 |

## 5. Discussion & Conclusions

The main topic addressed in this thesis is the automatic training of expressive voices. Several experiments were conducted on the acoustic domain, automatically selecting training data, and on semantic domain, predicting acoustics from semantics. In the last experiments, two state-of-the-art techniques were combined for the given task. Speech technology domain is an incredibly fast evolving field, especially being data driven where nowadays data is the key to all information technology. Nevertheless, the underlying technologies leveraged and presented in this work are not only not out of date, but are actually the driving power of current technologies. This includes the NN-based TTS and the semantic embeddings for text represention, but also not least i-vector-like representations for the acoustic domain. In that sense, this thesis is a substantial contribution to speech technology research.

## 6. Acknowledgements

I would like to thank especially Antonio Bonafonte, the tutor of this thesis. This work was supported by the Spanish Ministerio de Economía y Competitividad and European Regional Develpmend Fund, contract TEC2015-69266-P (MINECO/FEDER, UE) and by the FPU grant (Formación de Profesorado Universitario) from the Spanish Ministry of Science and Innovation (MCINN) to Igor Jauk, with a special collaboration with the University of Texas at El Paso, under Pr. Nigel Ward, and the National Institute of Informatics in Tokyo, under Pr. Junichi Yamagishi.

## 7. References

[1] R. Barra-Chicote, J. Yamagishi, S. King, J. Montero, and J. Macias-Guarasa. Analysis of statistical parametric and unit selection speech synthesis systems applied to emotional speech. *Speech Communication*, 52:394–404, 2010.

[2] Y. Bengio, R. Ducharme, P. Vincent, and C. Jauvin. A neural probabilistic model. *Journal of Machine Learning Research*, 3(Feb):1137–1155, 2003.

[3] P. Boersma and D Weenink. Praat: doing phonetics by computer (version 5.4.07), 2015.

[4] T. Bonafonte, P. Aguero, J. Adell, J. Perez, and A. Moreno. Ogmios: the UPC text-to-speech synthesis system for spoken translation. In *Proceedings of TC-STAR Workshop on Speech-to-Speech Translation*, pages 199–204, 2006.

[5] F. Burckhardt and W.F. Sendelmeier. Verification of acoustical correlates of emotional speech using formant synthesis. In *Proceedings of ISCA Workshop on Speech and Emotion*, pages 151–156, 2000.

[6] J.E. Cahn. Generation of affect in synthesized speech. In *Proceedings of American Voice I/O Society*, pages 251–256, 1989.

[7] L. Chen, M.J.F. Gales, N. Braunschweiler, M. Akamine, and K. Knill. Integrated expression prediction and speech synthesis from text. *IEEE Journal of Selected Topics in Signal Processing*, 8(2):323–335, 2014.

[8] N. Dehak, P.J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet. Front end factor analysis for speaker verification. *IEEE Transactions on Audio, Speech and Language Processing*, 19(4):788–798, 2011.

[9] D. Erro, I. Sainz, E. Navas, and I. Hernaez. Improved HNM-based vocoder for statistical synthesizers. In *Proceedings of Interspeech*, pages 1809–1812, 2011.

[10] F. Eyben. The opensmile book, 2016.

[11] F. Eyben, S. Buchholz, N. Braunschweiler, J. Latorre, V. Wan, M. Gales, and K. Knill. Unsupervised clustering of emotion and voice styles for expressive TTS. In *Proceedings of ICASSP*, pages 4009–4012, 2012.

[12] R.M. Gray. Vector quantization. *IEEE ASSP Magazine*, 1(2):4–29, 1984.

[13] W. Hamza, R. Bakis, E. Eide, M. Picheny, and J. Pitrelli. The IBM expressive speech synthesis system. In *Proceedings of ICSLP*, pages 2577–2580, 2004.

[14] I. Jauk. *Unsupervised Learning for Expressive Speech Synthesis*. PhD thesis, Universitat Politècnica de Catalunya, 2017.

[15] I. Jauk and A. Bonafonte. Direct expressive voice training based on semantic selection. In *Proceedings of Interspeech*, pages 3181–3185, 2016.

[16] I. Jauk and A. Bonafonte. Prosodic and spectral ivectors for expressive speech synthesis. In *Proceedings of Speech Synthesis Workshop 9*, pages 59–63, 2016.

[17] I. Jauk, A. Bonafonte, P. López-Otero, and L. Docio-Fernandez. Creating expressive synthetic voices by unsupervised clustering of audiobooks. In *Interspeech 2015*, pages 3380–3384, 2015.

[18] I. Jauk, J. Lorenzo-Trueba, J. Yamagishi, and A. Bonafonte. Expressive speech synthesis using sentiment embeddings. In *Proceedings of Interspeech*, pages 3062–3066, 2018.

[19] R. Kehrein. The prosody of authentic emotions. In *Proceedings of Speech Prosody*, pages 423–426, 2002.

[20] D. Klein and C.D. Manning. Accurate unlexicalized parsing. *ACL*, pages 423–430, 2003.

[21] K. Kuttler. *An introduction to linear algebra*. Bringham Young University, 2007.

[22] P. Lopez-Otero, L. Docio-Fernandez, and C. Garcia-Mateo. iVectors for continuous emotion recognition. In *Proceedings of Iberspeech 2014*, pages 31–40, 2014.

[23] J. Lorenzo Trueba. *Design and Evaluation of Statistical Parametric Techniques in Expressive Text-To-Speech: Emotion and Speaking Styles Transplantation*. PhD thesis, E.T.S.I. Telecomunicación (UPM), 2016.

[24] T. Masuko, K. Tokuda, T. Kobayashi, and S. Imai. Speech synthesis from hmms using dynamic features. In *Acoustics, Speech, and Signal Processing (ICASSP)*, pages 389–392, 1996.

[25] T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space. In *Proceedings of Workshop at ICLR*, 2013.

[26] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean. Distributed Representations of Words and Phrases and their Compositionality. *ArXiv e-prints*, October 2013.

[27] J.M. Montero, J. Gutierrez-Arriola, J. Colas, E. Enriquez, and J.M. Pardo. Analysis and modeling of emotional speech in spanish. In *Proceedings of ICPhS*, pages 671–674, 1999.

[28] I.R. Murray and J.L. Arnott. Toward the simulation of emotion in synthetic speech: a review of the literature on human vocal emotion. *The Journal of the Acoustic Society of America*, pages 1097–1108, 1993.

[29] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely. The kaldi speech recognition toolkit. In *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*. IEEE Signal Processing Society, December 2011. IEEE Catalog No.: CFP11SRW-USB.

[30] S. Reese, G. Boleda, L. Cuadros, M. Padró, and G. Rigau. Wikicorpus: A word-sense disambiguated multilingual wikipedia corpus. In *Proceedings of 7th Language Resources and Evaluation Conference (LREC10)*, pages 1418–1421, 2010.

[31] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn. Speaker verification using adapted Gaussian mixture models. *Digital Signal Processing*, 10(1-3):19–41, 2000.

[32] B. Schuller, R. Müller, M. Lang, and G. Rigoll. Speaker independent emotion recognition by early fusion of acoustic and linguistic features within ensembles. In *Proceedings of Interspeech*, pages 805–808, 2005.

[33] C.E. Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27(3):379–423, 1948.

[34] R. Socher, A. Perelygin, J.Y. Wu, J. Chuang, C.D. Manning, A.Y. Ng, and C. Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2013.

[35] E. Szekely, J. Cabral, P. Cahill, and J. Carson-Berndsen. Clustering expressive speech styles in audiobooks using glottal source parameters. In *Proceedings of Interspeech*, pages 2409–2412, 2011.

[36] S. Takaki and J. Yamagishi. Constructing a deep neural network based spectral model for statistical speech synthesis. *Recent Advances in Nonlinear Speech Processing*, 48:117–125, 2016.

[37] P. Wang, Y. Qian, F.K. Soong, L. He, and H. Zhao. Word embedding for recurrent neural network based tts synthesis. In *Proceedings of International conference on acoustics, speech and signal processing (ICASSP)*, pages 4879–4883, 2015.

[38] X. Wang, S. Takaki, and J. Yamagishi. Investigating of using continuous representation of various linguistic units in neural network based text-to-speech synthesis. *IEICE Transactions on Information and Systems*, E99-D(10):2471–2480, 2016.

[39] O. Watts. *Unsupervised Learning for Text-to-Speech Synthesis*. PhD thesis, University of Edinburgh, 2012.

[40] word2vec Tool for computing continuous distributed representations of words.

[41] J. Yamagishi, K. Onishi, T. Masuko, and T. Kobayashi. Modeling of various speaking styles and emotions for hmm-based speech synthesis. In *Proceedings of Eurospeech*, pages 2461–2464, 2003.

[42] Y. Zhao and G. Karypis. Empirical and theoretical comparisons of selected criterion functions for document clustering. *Machine Learning*, 55(3):311–331, 2004.