Acoustic Feature Prediction from Semantic Features for Expressive Speech using Deep Neural Networks

Igor Jauk, Antonio Bonafonte and Santiago Pascual
Universitat Politècnica de Catalunya
Barcelona, Spain
Email: {igor.jauk, antonio.bonafonte, santi.pascual}@upc.edu

***Abstract*—The goal of the study is to predict acoustic features of expressive speech from semantic vector space representations. Though a lot of successful work was invested in expressiveness analysis and prediction, the results often involve manual labeling, or indirect prediction evaluation such as speech synthesis. The proposed analysis aims at direct acoustic feature prediction and comparison to original acoustic features from an audiobook. The audiobook is mapped in a semantic vector space. A set of acoustic features is extracted from the same utterances, involving *iVectors* trained on *MFCC* and *F0* basis. Two regression models are trained with the semantic coordinates, DNNs and a baseline CART. Later, semantic and acoustic context features are combined for the prediction. The prediction is achieved successfully using the DNNs. A closer analysis shows that the prediction works best for larger utterances or utterances with specific contexts, and worst for general short utterances and proper names.**

I. Introduction

It seems obvious that there is expressiveness, i.e., emotions, attitudes, states of mind, moods, etc., codified in plain written text. If someone reads a book aloud, and somehow interprets the characters and the situations of the book, she or he will read them expressively, and normally, that expressiveness will be coherent to the people listening to the reader. The goal of this work is to determine whether expressiveness can be extracted from plain text and be represented as an acoustic term or descriptor. A book can have different characters, i.e., speakers, with a lot of different situations that have to be taken into account when analyzing expressiveness. Generalizing this claim it can be said that practically every discourse or conversation, no matter if in real life or as an interpretation, will codify information about the speaker, her/his emotions, social and cultural conditions, state of mind, state of health etc. It is very difficult to find a common term that would take into account all these things.

In general there have been different approaches to expression (and speaker) analysis in the literature: from the acoustic side, from the text side, and a mixture of both. For example [23] uses a set of 276 acoustic features to classify seven basic emotions from speech. A different approach is where emotions are not classified as discrete states, but rather points in a continuous space, for instance as suggested by [10]. In this framework [14] classifies emotions from speech using *i-vectors*. In [26] glottal source parameters are used as acoustic feature in *self organizing feature maps (SOFM)* to perform clustering of expressive speech styles in audiobooks. In [8], [9] clustering is also performed on audiobooks, but with a further step of creating synthetic voices from resulting clusters, in the latter case also i-vectors are integrated in the clustering.

On the other hand, text and linguistic information has been analyzed in order to determine its emotional content. For example, in [16] basic emotions are predicted from text using bag-of-words representations. In [25] different knowledge and corpus-based methods of emotion prediction from text are compared.

There are also studies that combine linguistic and acoustic features for emotions in different contexts. For instance, in [12], [13] emotions are classified in a call-center context using keywords and prosodic features. In [22] linguistic bag-of-words representations are combined with acoustic features to predict basic emotions. In [21] also bag-of-words representations are used on the text side; these are mapped to continuous emotion representations in a three-dimensional space, as proposed in [10]. For the evaluation a set of acoustic features is used to predict continuous emotion coordinates. In [3] no labels are predicted at all. Linguistic vector representations are directly mapped to *CAT* model weights, as described in [4].

In this work the goal is to analyze the expressive information encoded in plain text. As argued above, abstract labels are rather difficult to formulate since the expressive acoustic information is very variable. The goal is to predict from vector space representations of text directly acoustic parameters that are useful to represent expressiveness with its full variability. The acoustic features predicted here could be seen as an alternative to the continuous expressiveness representation. They have many more dimensions than three, but on the other hand the features do not have to be mapped nor labeled nor learned. They can be extracted directly from an acoustic database. Once predicted, they can be used for synthesis, but also for text and discourse analysis. If needed, they can also be mapped to labels.

The rest of the article is structured in the following way. Section II describes the general framework of the task and the databases used in the study. Section III presents the linguistic and acoustic features used for the prediction. Section IV presents several proposed regression models. Section V describes the experimental design and section VI shows the results of the experiments. Finally, section VII draws the final conclusion.

978-0-9928-6265-7/16/$31.00 ©2016 IEEE

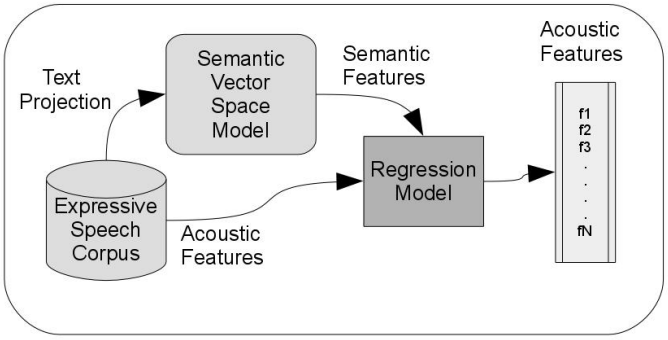

Fig. 1. Prediction system framework including acoustic features.

II. Framework

Figure 1 shows the general framework of the prediction method.

The first approach proposed in this work uses only linguistic features to predict acoustic feature vectors. The features used for the prediction are described in detail in section III.

In the first step, a *semantic vector space model (SVSM)* is constructed. Basically, the SVSM is a bag-of-words model represented as a continuous semantic space. It allows to project utterances into it and extract a set of coordinates for each utterance, 600 in this case. Section III describes in more detail the implementation of the SVSM used in this work.

The semantic vector representation reflects semantic knowledge of the text, where semantically similar concepts tend to occur close to each other in the space. This type of vector representation is used for text classification and information retrieval, for instance [1], [27] have used semantic vector representations for data selection for speech synthesis training.

Then, an expressive corpus is projected into the semantic vector space, such that each utterance of the semantic corpus can be described as a unique coordinate vector. At the same time, a set of acoustic features is extracted for each utterance from the expressive corpus that has been projected into the SVSM. In the next step, a regression model is trained to predict the acoustic features from the semantic coordinates. The prediction is designed to predict sets of utterances, such as paragraphs, so context information can be included in the predictor vectors.

The second framework includes past acoustic features in the predictor vectors. Each predictor vector includes, besides the semantic coordinates, the acoustic vector predicted for the previous utterance, i.e., the acoustic left context of the utterance. The first utterance has no left context, so a default acoustic vector is used as the left context of the first utterance. The default vector is extracted from a neutral phrase of the corpus.

A. Databases

The text database used for training of the SVSM is the Spanish portion of the *Wikicorpus*, containing 120 million words [19].

The acoustic database is an audiobook of 8.8 hours of duration, segmented on the sentence level. Some utterances that contained stuttering, reading errors, or noise imitations by the reader were removed, resulting in a total of 7903 utterances. The bad utterances have been identified partly by automatic tools and partly by manual revision. The segmentation was done using *Ogmios* speech analysis tools [2].

III. Features

Since the framework of the study relies on an audiobook database, there is a set of conditions that should be fulfilled in order to optimally codify the linguistic and the acoustic representations. Eventually, the expression prediction is carried out for large text instances, such as paragraphs. So the context of each utterance should be taken into account. The second point is that the audiobook in question, as books usually do, contains many characters, and although all characters are being imitated by the same reader, the ways how they express themselves are very different. For example, *anger* will probably be expressed very differently by a giant than by a witch. So we need acoustic features that would represent the different characters as different speakers.

A. Linguistic Features

The linguistic features are coordinates of corpus utterances mapped into the *semantic vector space model (SVSM)*. The SVSM is trained using the *skip-gram* method [15] implemented in the *word2vec* package [28], resulting in a 600 dimensional vector space. The number of dimensions has been determined experimentally to provide best results under acceptable training and execution time conditions, though surely there is space for improvement. One difference to most semantic vector space realization is that in this work the function words have not been removed. The decision is inspired by studies presented in [17] and tested in previous studies (unpublished) on semantic representations with and without function words, where best results were achieved including the stop words.

The linguistic feature vector is composed of three parts. The utterance in question, the left and the right contexts are projected into the SVSM and the coordinates are extracted, 1800 totally since the utterance and the context vectors have the length 600 each. The context on the left and on the right is composed of the next three words. The amount of words to take into account has been determined experimentally, the performance declining from the fourth word on. The reason might be that the context becomes too specific and moves away in the semantic space pushed by the words farther away from the sentence in question.

B. Acoustic Features

The acoustic features aim to represent expressiveness, not phonetic or segmental information, so each feature used in this work is suprasegmental, accounting for the whole sentence. Of course, the suprasegmental features, such as pitch, or rhythm, will also partly codify syntactic and other information.

Nevertheless, a previous study in [9] has shown that the feature set used in this work is useful to represent expressiveness.

Since the audiobook context implies not only the presence of different expressions, but also of different speakers, though only imitated by the reader, it is plausible to imply that acoustic features should account for the different speakers. A study conducted in [9] shows a significant performance improvement by including *i-vectors* as a feature for unsupervised clustering of an audiobook. Also in [14] i-vectors already have been used for emotion recognition.

i-vectors represent speech in a low-dimensional total variability subspace, which leads to a representation that is independent of the different sources of variability such as speaker, channel, noise, etc.

First, acoustic features are extracted from the waveform; in this work, 40 Mel-frequency cepstral coefficients and F0 values are used. Before extracting the i-vectors, a *Universal Background Model (UBM)* and the total variability matrix are trained as described in [20] and [11], respectively. In each case, the whole corpus was used for the training. The total variability matrix must be trained using audio segments that are homogeneous according to the speaker, channel and expressiveness. So silence was removed from the segments. Once the speech segments are obtained, Baum-Welch statistics are extracted using the UBM, which are used to obtain the total variability matrix that defines a total variability space, in which the speech segments are represented by a vector of total factors, namely i-vector [5].

Traditionally, i-vectors are calculated from MFCCs. Since prosody features are known to codify a significant amount of expressive information, in this work i-vectors are also calculated from F0. Additionally syllable and silence rates, means, variance and medians of durations are added to the acoustic vectors. In result, the acoustic feature vectors are composed of 600 dimensional i-vectors trained from MFCCs, 12 dimensional i-vectors trained from F0, and 8 dimensional vectors with syllable and silence statistics, 620 dimensions in total. The MFCCs and F0 features were extracted using *AHOCoder* [7]. The syllable and silence duration from the *Ogmios* speech analysis tools [2], and the i-vectors using the *Kaldi software* [18].

IV. Regression Models

Two different regression models were used to predict the acoustic feature vectors. The baseline model was the *Classification and Regression Trees (CART)*. A single tree was trained for each acoustic dimension.

Additionally, a *Deep Neural Network (DNN)* [6] was implemented to predict the feature vectors. The DNN is made of a stack of feed forward (*Dense*) layers, where each layer performs a projection followed by a nonlinearity, such that:

$$\mathbf{h} = g(\mathbf{W} \cdot \mathbf{x} + \mathbf{b}) \tag{1}$$

where $\mathbf{W}$ is the weights matrix, $\mathbf{x}$ is an input vector of features, $\mathbf{b}$ is the vector of biases and $\mathbf{g}$ is an element-wise non-linearity, which actually gives the DNN prediction capacity. There are several intermediate (hidden) layers, and in between, Dropouts [24] of 0.5 are applied to lower any possible over-fitting effect. At the output of the network a $tanh$ activation function is used, so the output features are normalized between $[-1, 1]$.

Figure 2 shows the general architecture of DNNs used in this work. After several experiments the best network design turns out to be a bottleneck design. Since the entrance layer has a rather larger number of neurons, the first hidden layer is also relatively large (1024 in the case of only semantic coordinates, 1500 for the semantic and acoustic combination). The next layer shrinks down to 256 neurons. There are several hidden layers with this number of neurons, which is then increased to 512, and to 620 in the output layer. In the case of the semantic prediction best results were achieved with 10 hidden layers. In the case of the semantic and acoustic combination the number of hidden layers is 5.

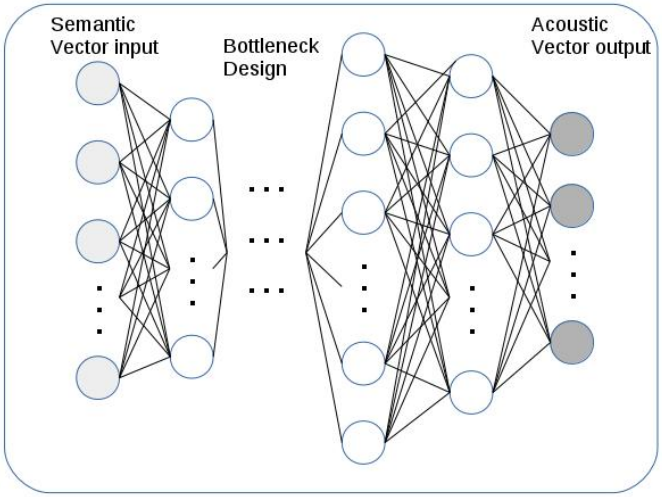

Fig. 2. DNN framework.

V. Experimental Design

For the experiment, first the correspondent models were trained. The systems which included the acoustic context were trained using the acoustic features of the previous utterances. In the prediction, the previously predicted feature vector was used for the next utterance. Four excerpts from the audiobook were selected for the evaluation, a total of 106 utterances. The test set was excluded from training. All test utterances and their context were projected into the SVSM obtaining the semantic coordinates. Then acoustic coordinates were predicted from the semantics for each of the four experimental conditions: (1) using CART with only semantics; (2) CART combined with acoustics; (3) DNNs with semantics only and (4) DNNs combined with acoustics.

The predicted acoustic feature vectors were compared to the original feature vectors for the utterances extracted from the corpus measuring the Euclidean distance, as in:

TABLE I
DISTANCE RESULTS. MEANS AND VARIANCES OF DISTANCES TO THE ORIGINAL ACOUSTIC FEATURE VECTORS.

	CART.sem	CART.acu	DNN.sem	DNN.acu	rand
MEAN	2.44	2.42	1.89	1.89	2.69
VAR	0.51	0.37	0.38	0.38	0.51

$$d(\mathbf{p},\mathbf{q}) = \sqrt{\sum_{i=1}^{n}(q_i - p_i)^2} \quad (2)$$

where **p** and **q** are the vectors to compare.

As a reference for the distance measure, the same test set was randomized and the distances were compared between the original and the shuffled vectors. Also, a closer analysis of distances between the original and the predicted vectors is conducted.

VI. RESULTS

Table I shows the results for the distance measures between the predicted feature vectors and the original feature vectors, in comparison to the distance of the original vectors to randomized original vectors. ANOVA is used to test the significance of the difference between these distances. Table II shows the ANOVA F-values for the distances.

Clearly, DNNs have produced predictions that significantly differ from random. CART did a slightly better prediction including the left acoustic context in the predictor vectors, reflected in the distance variance. However, looking at the ANOVA F values, although the F values are higher than the critical F value, the p values for the predictions with CART are 0.003 and 0.006, for the predictions with only semantic vectors and including the acoustics, respectively. So, the CART prediction is probably not significantly better in comparison to the shuffled data.

Between CARTs and DNNs, there is a significant difference in performance. However, combining semantic and acoustic features for the prediction did not result in any significant improvement.

TABLE II
ANOVA RESULTS BETWEEN THE FOUR CONDITIONS AND RANDOM. $\alpha = 0.05$, CRITICAL $F = 3.8861$. VALUES MARKED WITH * HAVE A p VALUE ABOVE 0.0025

	CART.sem	CART.acu	DNN.sem	DNN.acu
rand	7.852*	8.900*	76.897	76.908
CART.sem	-	0.044	43.551	43.561
CART.acu	-	-	40.520	40.530
DNN.sem	-	-	-	0.000

Figure 3 shows the distance plot of distances between the original acoustic vectors and the predicted vectors, for the four conditions, and the 106 utterances. The lower the line, the better is the prediction. The DNN predictions with semantics alone and with the combination with the acoustics are so similar that the lines practically overlap.

It can be observed that for some utterances the prediction is worse than for others. There are some peaks of larger distances, especially around the utterances 8, 36, 63 and in the area between 66 and 80. The utterance 8 is just a *"yes"*, so there is not much expressive information encoded, the utterance 36 is a proper name, also difficult to relate to prominent expressiveness, at least without taking into account larger context or world knowledge. The utterance 63 just says *"exclaims"*, with also very little context (*"perfect"* on the left and *"to bring your"* on the right). The area between 66 and 80 is a conversation in very general terms, including phrases like *"yes, please"*, *"hello"*, *"he said"* and some more.

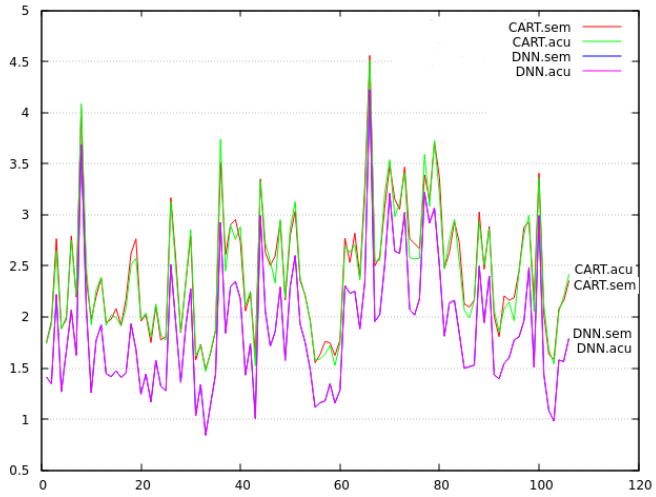

Fig. 3. Euclidean distance plot of the predicted to the original distances for the 106 utterances.

Possibly, rather large utterances codify more expressive information than short ones. Anyway, it is clear that reasonable prediction is truly possible for a reasonable portion of utterances, and that the deep neural networks show better performance for given task.

VII. SUMMARY AND CONCLUSIONS

The present study had the goal to analyze if expressiveness, in acoustic terms, can be predicted from plain text. For this task, each utterance of an audiobook was projected into a semantic vector space, and for each utterance three left and right context words, a 1800 dimensional set of coordinates was extracted. At the same time, for the same utterances acoustic feature vectors were computed. The feature vectors included i-vectors trained from MFCCs and from F0, in order to assure that speaker and expressive information is well represented in the feature vectors. Excluding a test set, two types of regression models were trained, using CART and DNNs. Each predictor model was trained with two conditions. First, only the semantic coordinates were included in the training process and second, the semantic coordinates were combined with the acoustic feature vector of the previous utterance, i.e., left

acoustic context. The prediction output is the acoustic feature vector for each of the test utterances. An objective evaluation was conducted to compare the predicted acoustic vectors to the original ones.

The results showed that the prediction using DNNs is far better than chance on this difficult task, however the CART model did not perform so well. Further, the combination of semantic vectors with the left acoustic context did not result in any significant improvement.

A more detailed analysis of the distances between the predicted and the original vectors showed that the prediction for some utterances worked better than for others. In particular, short sentences, where the semantics of the sentences had no clear expressive information, performed worst.

From given results it can be concluded that generally an expressiveness prediction from plain text is possible, although with limitations. Clearly, not all utterances codify expressive information. However, we as humans do know how to express them adequately in each situation. As seen from the analysis, the context is probably the clue point of how to provide the information to the system so the prediction can be improved. As argued above, blindly increasing the semantic context around the target utterance is not the solution, but possibly a finer training is needed is to find out which context is important in order to correctly predict the expressiveness, and which is not. Another interesting point, related to the context, if it is possible to include explicit world knowledge for a given domain and use it in the prediction.

ACKNOWLEDGMENT

This work was supported by the Spanish Ministerio de Economía y Competitividad and European Regional Developmend 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. A part of this work was realized during a research stay at the University of Texas at El Paso, also supported by the FPU grant. It also received funding from the Eusipco'11 organization.

REFERENCES

[1] F. Alas Pujol. *Conversión de texto en habla multidominio basada en seleccin de unidades con ajuste subjetivo de pesos y marcado robusto de pitch*. PhD thesis, Universitat Ramon Llull, 2006.

[2] A. 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.

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

[4] L. Chen, M.J.F. Gales, V. Wan, J. Latorre, and M. Akamine. Exploring rich expressive information from audiobook data using cluster adaptive training. In *Proceedings of Interspeech*, pages 958–961, 2012.

[5] 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, 2010.

[6] Li Deng and Dong Yu. Deep learning: Methods and applications. *Foundations and Trends in Signal Processing*, 7(3–4):197–387, 2014.

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

[8] 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.

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

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

[11] P. Kenny, G. Boulianne, and P. Dumouchel. Eigenvoice modeling with sparse training data. *IEEE Transactions on Speech and Audio Processing*, 13(3):345–354, 2005.

[12] Devillers L. and L. Lamel. Emotion detection in task-oriented dialogues. In *Proceedings of ICME 2003*, volume III, pages 549–552, 2003.

[13] C.M. Lee and R. Pieraccini. Combining acoustic and language information for emotion recognition. In *Proceedings of ICSLP 2002*, 2002.

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

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

[16] S. Ovesdotter Alm, D. Roth, and R Sproat. Emotion from text: machine learning for text-based emotion prediction. In *Proceedings of Conf. HLT-EMNLP*, pages 579–586, 2005.

[17] J.W. Pennebaker. *The Secret Life of Pronouns*. 2011.

[18] Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, Jan Silovsky, Georg Stemmer, and Karel 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.

[19] S. Reese, G. Boleda, L. Cuadros, M. Padr, 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.

[20] 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.

[21] B. Schuller. Recognizing affect from linguistic information in 3d continuous space. *IEEE Transactions on Affective computing*, 2(4):192–205, 2000.

[22] B. Schuller, R. Mller, 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.

[23] B. Schuller, S. Reiter, R. Muller, M. Al-Hames, M. Lang, and G. Rigoll. Speaker independent speech emotion recognition by ensemble classification. In *Proceedings of IEEE int. conf. multimedia and expo*, pages 864–867, 2005.

[24] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 15(1):1929–1958, 2014.

[25] C. Strapparava and R. Mihalcea. Learning to identify emotions in text. In *Proceedings of 2008 ACM Symposium on Applied Computing*, pages 1556–1560, 2008.

[26] 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.

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

[28] word2vec. Tool for computing continuous distributed representations of words. http://www.gnu.org/software/gsl/.