INTERSPEECH 2019
September 15–19, 2019, Graz, Austria

Prosodic Representations of Prominence Classification Neural Networks and Autoencoders Using Bottleneck Features

Sofoklis Kakouros[1], *Antti Suni*[1], *Juraj Šimko*[1], *Martti Vainio*[1]

[1]Department of Digital Humanities, University of Helsinki, Finland

{sofoklis.kakouros,antti.suni,juraj.simko,martti.vainio}@helsinki.fi

Abstract

Prominence perception has been known to correlate with a complex interplay of the acoustic features of energy, fundamental frequency, spectral tilt, and duration. The contribution and importance of each of these features in distinguishing between prominent and non-prominent units in speech is not always easy to determine, and more so, the prosodic representations that humans and automatic classifiers learn have been difficult to interpret. This work focuses on examining the acoustic prosodic representations that binary prominence classification neural networks and autoencoders learn for prominence. We investigate the complex features learned at different layers of the network as well as the 10-dimensional bottleneck features (BNFs), for the standard acoustic prosodic correlates of prominence separately and in combination. We analyze and visualize the BNFs obtained from the prominence classification neural networks as well as their network activations. The experiments are conducted on a corpus of Dutch continuous speech with manually annotated prominence labels. Our results show that the prosodic representations obtained from the BNFs and higher-dimensional non-BNFs provide good separation of the two prominence categories, with, however, different partitioning of the BNF space for the distinct features, and the best overall separation obtained for F0.

Index Terms: prosody, prominence, neural networks, autoencoder, bottleneck features, prominence classification

1. Introduction

Spoken language contains multiple levels of information, ranging from linguistic content to cues about the speaker. Prosody can be broadly seen as a level of representation that reflects acoustic-phonetic variation that extends across long temporal segments in speech and conveys information in addition to the lexical content. In general, prosody and prosodic phenomena involve aspects of speech that extend the individual phoneme and may cover sequences of words and entire phrases (see [1, 2, 3] for related definitions). Prominence is a prosodic phenomenon that conveys the subjective impression of emphasis and is defined as the perception of a linguistic unit standing out from its environment (see [4, 5, 6] for related definitions). Earlier, many studies focused on determining the acoustic correlates of prominence [7, 8, 9], and, more recently, on methods for its automatic detection [10, 11, 12, 13, 14]. One interesting aspect on the study of prominence that has been enabled by the success of deep neural networks (DNNs), and that has not been widely explored, is whether DNNs are capable of learning prominence-like representations of speech. In particular, understanding the learning behavior and internal representations of prominence by DNNs can potentially provide interesting and important insights regarding the acoustic prosodic characterization of prominence.

Earlier work on prominence has established a number of different features that seem to hold a role in the acoustic characterization of prominent units in speech. In particular, four acoustic features have been found to correlate with the incidence of prominent units in speech: energy [9, 15, 7], fundamental frequency [8], spectral tilt [16, 17], and duration [15, 7]. In general, there seems to be a complex interplay of the four acoustic correlates of prominence where the exact acoustic specification of prominence-encoding features cannot be always easily determined (see, e.g., [7]). This becomes particularly evident when considering that different feature or feature set specifications may be descriptive of prominence (e.g., [18]) but also when looking into the large inter-annotator differences in marking prominences [10]. The latter observation makes the study of the complex feature representations that DNNs learn over the acoustic prosodic space particularly interesting as they can potentially point to the aspects of the acoustic space that are most helpful in identifying prominence categories.

Previous work on exploring how neural networks represent different aspects of speech has focused primarily on investigating the learning of phonetic representations [19, 20, 21]. This typically involves two general approaches: (i) an unsupervised one, where a neural network autoencoder is used in order to investigate whether the network is capable of learning phoneme-like representations without explicit labels [22], (ii) a supervised one, where the neural network is trained with phone labels for the task of phoneme recognition [20, 19]. In both cases, investigations are focused on analyzing the representational properties of the complex features learned at different layers of the network and also at different nodes. In the case of the autoencoder the interest also falls in how and whether the reduced (compressed) representations of the neural network at a bottleneck (a layer having a smaller number of hidden units compared to other hidden layers) can represent the phonetic categories (e.g., [22]). In general, this type of bottleneck features (BNFs) have been shown to be effective in learning low-dimensional representations of high-dimensional inputs [23].

In this work we investigate the acoustic prosodic representations that neural networks learn when distinguishing between prominent and non-prominent units in speech. As prosody and prosodic phenomena reflect variations that extend beyond single phonetic segments, our aim is to explore large acoustic contexts over the three prosodic correlates of prominence: energy, F0, and spectral tilt. The experimental design involves the use of generic feed-forward DNNs on two tasks: (i) a binary prominence classification task, and (ii) an autoencoder DNN that compresses the acoustic input to 10-dimensionsal BNFs. In addition, we run standard supervised classification on our data to obtain a baseline result of the prominence class separation. Finally, we also visualize our results in order to get better insights of the network representations at the different hidden layers and observe the distinct patterns that are formed. Our experiments

http://dx.doi.org/10.21437/Interspeech.2019-2984

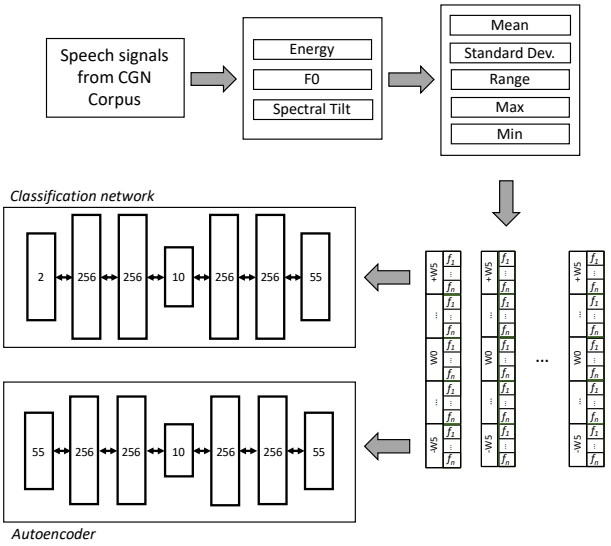

Figure 1: *Overview of the experimental setup.*

are conducted on a corpus of continuous Dutch speech with manual annotations of prominence.

2. Materials and Methods

The material used in this work consists of Dutch continuous speech taken from Dutch news broadcast recordings. Data are analyzed using the three acoustic prosodic features of energy, F0, and spectral tilt. All features are further processed into vectors that are then fed to two distinct neural networks (see also Fig. 1). These are further described next.

2.1. The Spoken Dutch Corpus

In the current experiments, the Spoken Dutch Corpus (Corpus Gesproken Nederlands; CGN) was used in order to evaluate the prominence class differences in the DNNs for Dutch continuous speech [24]. The CGN corpus is a database of contemporary standard Dutch containing material spoken by adults in the Netherlands and Flanders. CGN contains nearly 9 million words that correspond to approximately 800 hours of speech data. Two thirds of the CGN material consist of recordings from the Netherlands and one third from Flanders. The corpus contains various manually generated or verified annotations such as phonetic transcriptions, word level alignment, and prosodic annotations (see [25] for a description). For the current study, the prosodically annotated subset of the Dutch news broadcast (*component k*) section of the corpus was utilized that consists of 134 news broadcasts spoken by 10 different speakers (9 male and 1 female) and contains a total of 7438 words (44.3 minutes of speech data; 42.6% of the words marked as prominent with the average per speaker ratio of prominent words being 0.43 with σ=0.03). Each sentence of the prosodically annotated subset of *component k* was hand-labeled by two trained annotators (see [26] for a description).

For the classification experiments, a 10-fold cross-validation procedure was used where, at each fold, one speaker was left for testing and the remaining 9 were used for training the classification network —no same speaker occuring at the same time in the training and test sets.

2.2. Prosodic Features

2.2.1. Feature extraction

Energy, F0, and spectral tilt were used as the primary features in the experiments. Word durations were extracted from the corpus annotations but were not added explicitly into the data (word durations are used for the computation of the feature descriptors over words). Speech signals were first downsampled to 8 kHz and all features were computed using a 25-ms window 5-ms frame shift. F0 contours were computed using the YAAPT pitch tracking algorithm [27], spectral tilt was computed from the Mel-frequency cepstral coefficients (MFCCs) and taking the first (C1) MFCC (see, e.g., [28, 29]), and signal energy based on Eq. (1) (where x denotes the signal, *t* the current sample, τ the frame shift, and *w* the frame length; see, e.g., [10]).

$$EN(n) = \sum_{\tau=-\frac{w}{2}}^{\tau=\frac{w}{2}-1} |x(t+\tau)|^2 \quad (1)$$

2.2.2. Normalization

All computed acoustic features were normalized in order to account for inter- and intra-talker variation. In particular, F0 was semitone normalized with respect to the median F0 for each speaker according to Eq. (2), spectral tilt was z-score normalized per speaker, and energy was logarithmically normalized.

$$F0'(n) = 12 \cdot \log_2 \left(\frac{F0(n)}{F0_{median}} \right) \quad (2)$$

2.2.3. Statistical descriptors and word-vectors

To obtain prosodic representations over larger contexts, words were selected as the unit of analysis and five statistical descriptors were computed using the normalized feature values. In particular, the five descriptors utilized are the: *mean*, *standard deviation*, *maximum*, *minimum*, and *range* (the difference between the maximum and minimum for a specific feature over a word). For each word and each feature, one vector was constructed containing the five descriptors.

After the word-level computation of the statistical descriptors, and in order to represent larger acoustic contexts in the data, for each word, the preceding and forthcoming five word vectors were included in each center word (see also also Fig. 1). This resulted to 55-dimensional vectors for each word in the data. Note also that in the case of feature combinations, the resulting vector dimensions were 110 and 165 when two and three features were combined respectively —for simplicity of presentation we refer to the basic setup using one feature and 55-dimensional vectors.

2.3. Prominence classification neural network

For the prominence classification task we built a standard feed-forward neural network with densely connected layers (see also Fig. 1). The network input was 55-dimensional word vectors and the output a 1-dimensional binary prominence class label. Inputs to the network were z-scored normalized across all data to ensure proper scaling. The network was configured using rectified linear unit activation functions for the hidden layers, a sigmoid output layer, 100 epochs, minibatch size of 50, and a configuration layout of dimensions d = [256,256,10,256,256] for the hidden layers. The model was trained using Adam optimizer with a learning rate of 0.001 and with binary cross-entropy as

Table 1: *Prominence classification performance for all acoustic features and their combinations for the CGN component-k data. Values in bold indicate the best feature and feature combination performance.*

	ACC	PRC	RCL	F
EN	77.12	77.23	77.12	77.10
F0	**77.53**	**77.73**	**77.73**	**77.43**
ST	75.50	75.56	75.50	75.53
EN+F0	81.73	81.81	81.73	81.70
EN+ST	77.91	78.11	77.91	77.80
F0+ST	81.30	81.48	81.30	81.24
EN+F0+ST	**82.50**	**82.64**	**82.51**	**82.49**

the cost function.

2.4. Autoencoder

For the autoencoder, similarly as in the classification network, a standard feed-forward neural network with densely connected layers was used. The network input and output was 55-dimensional word vectors. Inputs to the network were z-scored normalized across all data. The network was configured using rectified linear unit activation functions for the hidden layers, 100 epochs, minibatch size of 50, and a configuration layout of dimensions d = [256,256,10,256,256] for the hidden layers. The model was trained using Adam optimizer with a learning rate of 0.001 and with binary mean squared error (MSE) as the cost function.

2.5. Visualization

To visualize the complex feature representations at the different hidden layers we needed to reduce the high-dimensional data to two-dimensional representations. For this purpose, we utilized standard principal component analysis (PCA) over the hidden network activations.

2.6. Evaluation

To evaluate the baseline classification performance in our data we compared the manual binary prominence markings from the corpus annotations with the word-level hypotheses provided by the classifier. As an additional reference, in some of the reported experiments, we also used the gender and speaker labels from the corpus annotations. To measure performance, we used the model evaluation metrics of precision (PRC), recall (RCL), their harmonic mean (F-value), as well as accuracy (ACC).

3. Results

We first present the results of the supervised classification experiments for prominence and then the experiments on the acoustic prosodic representations of the two neural networks.

3.1. Supervised classification

The prominence classification neural network (see section 2.3) was evaluated for the features of energy, F0, spectral tilt, and their combinations in a 10-fold cross-validation setup. The results are presented in Table 1. The best individual feature performance in our data was reached for F0 with an accuracy of 77.53% and the best feature combination for energy, F0, and spectral tilt with 82.5%. Overall, all features seem to contribute in the classification of prominence, with a combination of all features giving the best performance in the task.

Table 2: *kNN performance of the representations across the hidden layers for the classification network and the autoencoder. Results are presented with reference to prominence, gender, and speaker labels.*

		Hidden Layer					
		256	**256**	**10**	**256**	**256**	
Prom.	Classif. Net.	80.49	83.68	88.35	88.61	88.61	F0
	AE	70.77	67.52	64.57	65.6	67.10	
Speaker	Classif. Net.	40.01	36.88	28.72	28.19	26.55	
	AE	40.34	38.52	35.72	35.36	35.91	
Gender	Classif. Net.	93.44	91.69	89.69	89.68	89.34	
	AE	97.66	97.47	97.80	96.90	96.95	
		256	**256**	**10**	**256**	**256**	
Prom.	Classif. Net.	78.73	82.92	87.46	87.60	87.47	ST
	AE	66.44	63.27	62.32	61.80	61.41	
Speaker	Classif. Net.	33.69	30.59	25.79	25.87	25.18	
	AE	22.97	21.52	20.67	19.75	20.15	
Gender	Classif. Net.	90.75	89.67	89.36	89.32	89.42	
	AE	89.73	89.53	88.95	89.63	89.54	
		256	**256**	**10**	**256**	**256**	
Prom.	Classif. Net.	79.45	84.63	88.38	88.39	88.39	EN
	AE	66.15	61.45	58.85	59.71	59.71	
Speaker	Classif. Net.	36.31	33.83	26.37	26.71	25.53	
	AE	29.33	27.04	26.78	25.67	26.15	
Gender	Classif. Net.	90.46	89.86	89.72	89.40	89.58	
	AE	90.40	90.37	90.44	90.04	89.86	

3.2. Prominence classification representations

The prosodic representations that binary prominence classification networks learn were investigated for all features and their combinations (see Table 2; only three individual acoustic features are presented here due to space limitations). To quantify the differences in the hidden layers network activations for the distinct features and their combinations, we used the *k*-nearest neighbor (kNN) algorithm. kNN was given the layer activations as input (separately for each layer) with three different sets of labels as reference: the prominence class, gender, and speaker label. The inclusion of the additional labels was aimed at gaining a better understanding of the different aspects of speech that the hidden layers might be representing. kNN was run on randomly selected test activations (10-fold setup) combined with equal number of randomly selected training activations.

The results for the classifier indicate, as expected, that the hidden layers are becoming increasingly better at discriminating between the two prominence classes (see Table 2). Interestingly, the classifier already from the first hidden layer seems to be capable of performing a good discrimination of the gender of the speakers and less so, of the speakers' label. This discriminatory capacity for both the gender and speaker is decreasing with the successive hidden layers towards the network output. In general, the prosodic representations for prominence seem to be relatively robust across the hidden layers for all features examined.

3.3. Autoencoder representations

In contrast to the prosodic representations of the prominence classification network, the autoencoder representations create a

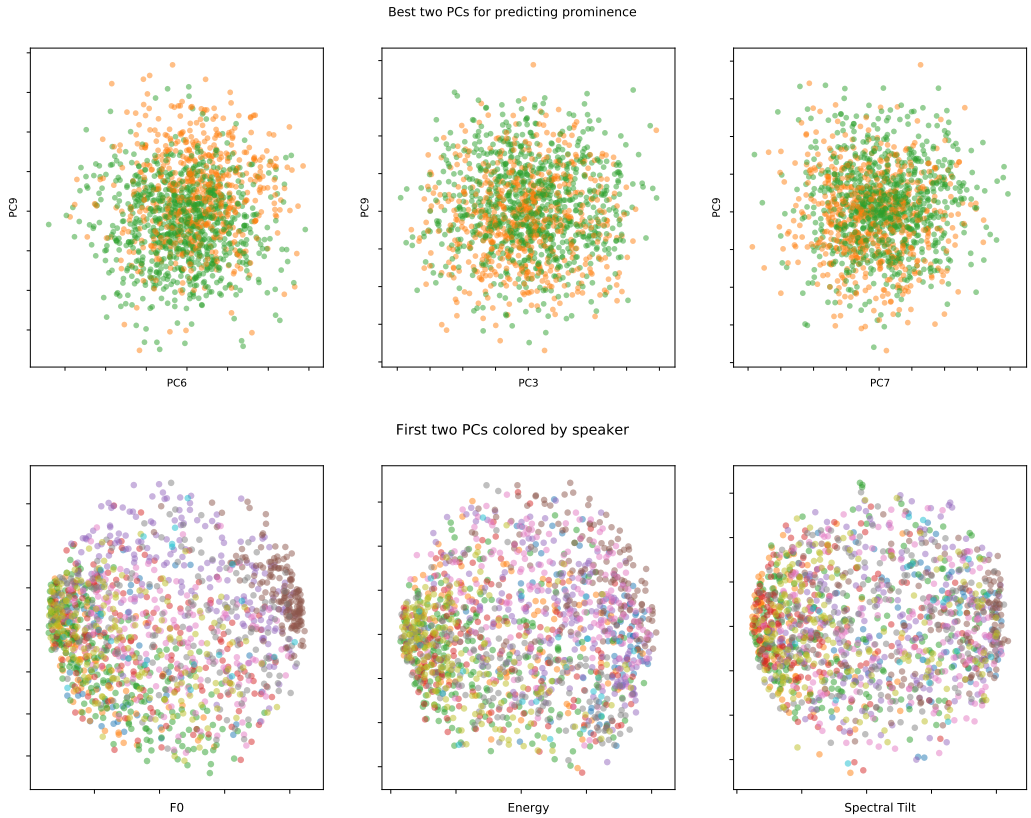

Figure 2: *Visualization of the autoencoder bottleneck code for F0, energy, and spectral tilt. Top panel: plots of the two principal components that best describe prominence where yellow denotes the prominent and green the non-prominent classes. Bottom panel: plots of the first and second principal components describing speakers.*

different overall picture —for the autoencoder, the same type of analysis was performed using kNN and the results are presented in Table 2. In this case, there are clear differences between the distinct features in how well they can discriminate between the prominence classes. In particular, the feature with the best overall prosodic representation for prominence is F0, whereas the different feature combinations do not seem to improve over F0. Interestingly, the autoencoder seems to be able to discriminate well the prominence classes while performing the best separation for the gender and the worst for speakers' labels. It is noteworthy that although the speakers' performance is low for the autoencoder, it is higher than the best performance on the same labels the classification network reached. These patterns are preserved in the network and are observed also in the bottleneck layer of the network. Fig. 2 presents an overview of the main findings for the 10-dimensional BNFs of the autoencoder.

4. Discussion and Conclusions

The results presented in this work provide insights on the prosodic representations that neural networks learn when evaluating prominence. In particular, it was shown that prosodic representations obtained from the BNFs and higher-dimensional non-BNFs provide good separation of the two prominence categories for both the binary prominence classification task and the autoencoder. It seems that the autoencoder is capable of representing and preserving the separation of the prominence categories across the layers, but more so, it can also discriminate between other acoustic aspects in the signal such as the gender and speaker information.

Despite normalization of the acoustic features, principal component analysis of the autoencoder bottleneck features (see Fig. 2) suggests that much of the variation observed in the acoustic features is due to speaker specific idiosyncrasies. Observing the speaker performance results in Table 2, it appears that the classification network learns to normalize this variance in early layers, while the top layers concentrate on the actual task of prominence classification.

Furthermore, the results from the binary prominence classification provided additional evidence of the importance of the three acoustic correlates of prominence for Dutch. These results are also close to those of an earlier study on the same data [13].

In this exploratory study we investigated the role of three acoustic features over words within a fixed temporal (word) context. In future work, we aim to extend the experiments with larger datasets and also include additional features, such as the Mel-frequency cepstral coefficients (MFCCs) and duration. In addition, it would be of interest to investigate the importance of using different underlying linguistic units, such as syllables, and different ways of modeling the temporal context.

5. Acknowledgements

This study was partly funded by the Academy of Finland project *Digital Language Typology: Mining from the Surface to the Core* (project no. 12933481).

6. References

[1] S. Shattuck-Hufnagel and A. E. Turk, "A prosody tutorial for investigators of auditory sentence processing," *Journal of psycholinguistic research*, vol. 25, no. 2, pp. 193–247, 1996.

[2] A. Cutler, D. Dahan, and W. Van Donselaar, "Prosody in the comprehension of spoken language: A literature review," *Language and speech*, vol. 40, no. 2, pp. 141–201, 1997.

[3] M. Wagner and D. G. Watson, "Experimental and theoretical advances in prosody: A review," *Language and cognitive processes*, vol. 25, no. 7-9, pp. 905–945, 2010.

[4] J. Terken and D. Hermes, "The perception of prosodic prominence," in *Prosody: Theory and experiment*, H. M., Ed. Springer, 2000, pp. 89–127.

[5] P. Wagner, A. Origlia, C. Avesani, G. Christodoulides, F. Cutugno, M. D'Imperio, D. E. Mancebo, B. G. Fivela, A. Lacharet, B. Ludusan *et al.*, "Different parts of the same elephant: a roadmap to disentangle and connect different perspectives of prosodic prominence," in *International Congress of Phonetic Sciences (ICPhS 2015)*. International Phonetic Association, 2015.

[6] A. Cutler, "Lexical stress," in *The handbook of speech perception*, D. B. Pisoni and R. E. Remez, Eds. Oxford: Blackwell, 2005, pp. 264–289.

[7] P. Lieberman, "Some acoustic correlates of word stress in american english," *The Journal of the Acoustical Society of America*, vol. 32, no. 4, pp. 451–454, 1960.

[8] J. Terken, "Fundamental frequency and perceived prominence of accented syllables," *The Journal of the Acoustical Society of America*, vol. 89, no. 4, pp. 1768–1776, 1991.

[9] G. Kochanski, E. Grabe, J. Coleman, and B. Rosner, "Loudness predicts prominence: Fundamental frequency lends little," *The Journal of the Acoustical Society of America*, vol. 118, no. 2, pp. 1038–1054, 2005.

[10] S. Kakouros and O. Räsänen, "3pro–an unsupervised method for the automatic detection of sentence prominence in speech," *Speech Communication*, vol. 82, pp. 67–84, 2016.

[11] F. Tamburini, C. Bertini, and P. M. Bertinetto, "Prosodic prominence detection in italian continuous speech using probabilistic graphical models," in *Proceedings of Speech Prosody*, 2014, pp. 285–289.

[12] G. Christodoulides and M. Avanzi, "An evaluation of machine learning methods for prominence detection in french," in *Proceedings of the Fifteenth annual conference of the International Speech Communication Association*. International Speech Communications Association, 2014, pp. 116–119.

[13] S. Kakouros, J. Pelemans, L. Verwimp, P. Wambacq, and O. Räsänen, "Analyzing the contribution of top-down lexical and bottom-up acoustic cues in the detection of sentence prominence." in *Proceedings of Interspeech*. International Speech Communications Association, 2016, pp. 1074–1078.

[14] A. Suni, J. Šimko, D. Aalto, and M. Vainio, "Hierarchical representation and estimation of prosody using continuous wavelet transform," *Computer Speech & Language*, vol. 45, pp. 123–136, 2017.

[15] D. B. Fry, "Duration and intensity as physical correlates of linguistic stress," *The Journal of the Acoustical Society of America*, vol. 27, no. 4, pp. 765–768, 1955.

[16] A. M. Sluijter and V. J. Van Heuven, "Spectral balance as an acoustic correlate of linguistic stress," *The Journal of the Acoustical society of America*, vol. 100, no. 4, pp. 2471–2485, 1996.

[17] N. Campbell and M. Beckman, "Stress, prominence, and spectral tilt," in *Intonation: Theory, Models, and Applications (Proceedings of an ESCA Workshop)*, A. Botinis, G. Kouroupetroglou, and G. Carayiannis, Eds., 1997, pp. 67–70.

[18] C. Gussenhoven, B. H. Repp, A. Rietveld, H. H. Rump, and J. Terken, "The perceptual prominence of fundamental frequency peaks," *The Journal of the Acoustical Society of America*, vol. 102, no. 5, pp. 3009–3022, 1997.

[19] T. Nagamine, M. L. Seltzer, and N. Mesgarani, "Exploring how deep neural networks form phonemic categories," in *Proceedings of Interspeech*. International Speech Communications Association, 2015, pp. 1912–1916.

[20] L. Bai, P. Weber, P. Jancovic, and M. Russell, "Exploring how phone classification neural networks learn phonetic information by visualising and interpreting bottleneck features." International Speech Communications Association, 2018, pp. 1472–1476.

[21] O. Scharenborg, N. van der Gouw, M. Larson, and E. Marchiori, "The representation of speech in deep neural networks," in *Proceedings of the International Conference on Multimedia Modeling*. Springer, 2019, pp. 194–205.

[22] O. Räsänen, T. Nagamine, and N. Mesgarani, "Analyzing distributional learning of phonemic categories in unsupervised deep neural networks," in *Proceedings of the Annual Conference of the Cognitive Science Society*. Cognitive Science Society (US), 2016, pp. 1757–1762.

[23] L. Deng, M. L. Seltzer, D. Yu, A. Acero, A.-r. Mohamed, and G. Hinton, "Binary coding of speech spectrograms using a deep auto-encoder," in *Proceedings of the Annual Conference of the International Speech Communication Association*. International Speech Communications Association, 2010, pp. 1692–1695.

[24] N. Oostdijk, W. Goedertier, F. v. Eynde, L. Boves, J.-P. Martens, M. Moortgat, and R. H. Baayen, "Experiences from the spoken dutch corpus project," in *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC-2002), Las Palmas, Gran Canaria*. ELRA, 2002, pp. 340–347.

[25] J. Duchateau, T. Ceyssens, and H. Van Hamme, "Use and evaluation of prosodic annotations in dutch," in *Proceedings of the International Conference on Language Resources and Evaluation (LREC-2004), Lisbon, Portugal*. ELRA, 2004, pp. 1517–1520.

[26] J. Buhmann, J. Caspers, V. J. van Heuven, H. Hoekstra, J.-P. Martens, and M. Swerts, "Annotation of prominent words, prosodic boundaries and segmental lengthening by non-expert transcribers in the spoken dutch corpus." in *Proceedings of the International Conference on Language Resources and Evaluation (LREC-2002), Las Palmas, Gran Canaria*. ELRA, 2002, pp. 779–785.

[27] S. A. Zahorian and H. Hu, "A spectral/temporal method for robust fundamental frequency tracking," *The Journal of the Acoustical Society of America*, vol. 123, no. 6, pp. 4559–4571, 2008.

[28] S. Kakouros, O. Räsänen, and P. Alku, "Comparison of spectral tilt measures for sentence prominence in speecheffects of dimensionality and adverse noise conditions," *Speech Communication*, vol. 103, pp. 11–26, 2018.

[29] S. Kakouros, O. Räsanen, and P. Alku, "Evaluation of spectral tilt measures for sentence prominence under different noise conditions." in *Proceedings of Interspeech*, 2017, pp. 3211–3215.