10th International Conference on Speech Prosody 2020
25-28 May 2020, Tokyo, Japan

Analysis of speech prosody using WaveNet embeddings: The Lombard effect

Juraj Šimko, Martti Vainio, Antti Suni

University of Helsinki, Finland
juraj.simko@helsinki.fi

Abstract

We present a novel methodology for speech prosody research based on the analysis of embeddings used to condition a convolutional WaveNet speech synthesis system. The methodology is evaluated using a corpus of Lombard speech, pre-processed in order to preserve only prosodic characteristics of the original recordings. The conditioning embeddings are trained to represent the combined influences of three sources of prosodic variation present in the corpus: the level and type of ambient noise, and the sentence focus type. We show that the resulting representations can be used to quantify the prosodic effects of the underlying influences, as well as interactions among them, in a statistically robust way. Comparing the results of our analysis with the results of a more traditional examination indicates that the presented methodology can be used as an alternative method of phonetic analysis of prosodic phenomena.

Index Terms: WaveNet, embeddings, Lombard speech, sentence focus, noise type

1. Introduction

Many modern machine learning systems learn to map a set of parameters (a specification of the given task) to output that satisfies the requirements encoded in the input parameters. Often, such a system essentially learns a (potentially very complex) statistical model of the training data in a form of conditional probabilities of output patterns given the input parameters. In some cases, the system's architecture includes modules dedicated to "translating" some of the input parameters – some of them expressed in a categorical format with arbitrary labels – to a continuous numerical form that is subsequently processed by the rest of the system.

In many implementations of deep network speech synthesis architecture, the information determining what kind of signal the network generates (what segment, in what voice, in what style), is fed to the system by a densely connected conditioning layer. This conditioning is standardly implemented in a form of embedding layer that learns, in parallel to the rest of the network, to map each categorical label to a numerical vector of activations [1, 2].

Presumably, the more similar the generated outputs conditioned by different categorical inputs, the more similar the conditioning vectors encoding the inputs. As has recently been shown using a corpus of dialectal variation in Swedish [3], the analysis of the embedding vector space may indeed reveal typologically relevant information about the relationships among conditions (dialects) in terms of their effects on speech prosody.

The aim of the current work is to present and significantly extend the methodology of prosodic investigation based on the analysis of embedding vector spaces. We show that the methodology is compatible with the more traditional phonetic approaches that use a limited range of signal properties extracted from speech material, such as particular f_0 or intensity values.

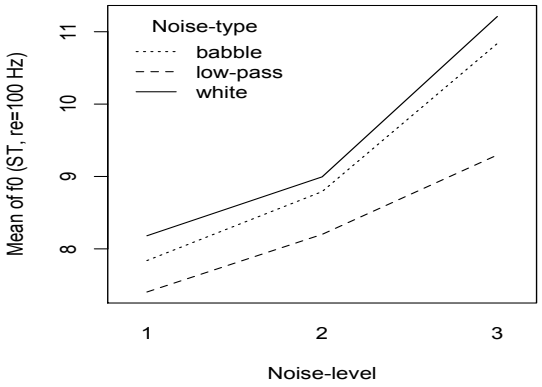

Figure 1: *Mean f_0 level vs. noise-level (1 = 60 dB, 2 = 70 dB, 3 = 80 dB; no-noise = 6.63 ST), from [4].*

In particular, we use the methodology for a prosodic re-analysis of Lombard speech material previously analyzed in [4].

Lombard effect is a set of adaptations of speech in a loud environment [5]. The adaptations include intensity and fundamental frequency (f_0) increase, changes of voice quality and durational characteristics of speech as well as changes of spectral properties of segments and is known to be influenced by the ambient noise loudness and type but also linguistic content and speaker characteristics [6, 7, 8, 9].

The Lombard speech corpus investigated here (and in [4]) contains multiple renditions of Finnish sentences varying in several dimensions, namely in sentence *focus* type as well as *loudness level* and *type* of ambient noise presented to the speakers during recording sessions. As illustrated in Fig. 1, the original phonetic analysis showed that the mean f_0 over the utterances increased with the increasing loudness of the ambient noise and that the f_0 increase was systematically affected by the noise type.

The same corpus is used here to train a WaveNet-based synthesis system conditioned by a set of labels comprising all possible combinations of the noise loudness level, type and focus condition as well as by a separate conditioning on speaker's identity. We extracted the conditioning embeddings for multiple trained instances of the system. Subsequently, we statistically analysed the topology of all these embedding spaces in terms of relationships among categories encoded in the labels. The results of the analysis are compatible with and extend the findings of the original phonetic analysis of the corpus.

2. WaveNet-based prosodic analysis

A globally conditioned WaveNet-based synthesis system was trained on a corpus of Lombard speech data with material containing 12 Finnish sentences uttered in three different focus patterns in presence of ambient noise.

10.21437/SpeechProsody.2020-186

2.1. Material

2.1.1. Corpus

The speech material used in this work is identical to the corpus analysed in [4]. The corpus contains recordings of 21 native Finnish speakers uttering 12 Finnish sentences. All sentences have subject–verb–object structure, and are matched in terms of number of syllables as well as phonological quantity patterns. Three different types of focus are elicited for each sentence: broad focus, narrow focus on the sentence subject (N1) and narrow focus on the object (N2). The utterances were recorded in silent environment as well as in the presence of ambient noise of 60 dB, 70 dB and 80 dB SPL, played through closed headphones. Three types of the ambient noise were used (scaled for appropriate loudness): babble noise from the NOISEX-92 database [10], white noise and a low-pass filtered noise with cut-off frequency of 1 kHz (see [4] for details).

The space of all possible combinations (7560) is sparsely sampled in the database in a balanced way; the corpus contains 2520 utterances in total, i.e., each speaker read four out of 12 sentences in each noise-type/noise-level/focus-type combination.

The entire corpus is manually annotated at syllable level by trained phoneticians.

2.1.2. Prosodic signal

In the present work we aim at the analysis of the Lombard speech material in terms of its prosody rather than segmental information. Therefore, low sample rate "prosodic" signals were created from the original speech waveforms, preserving only the main prosodic characteristics of the signal, namely f_0 and energy envelope, in the following way.

First, f_0 contours (in Hz) were extracted from the speech material; interpolated contours were used for unvoiced intervals. Then, the f_0 contours were used to generate an envelope-modulated sinusoidal signal:

$$s(t) = e(t) \sin\big(2\pi \int_0^t f_0(\tau) d\tau\big),$$

where $e(t)$ is the energy envelope of the original waveform; $s(t)$ has the same f_0 and energy envelope as the original, but contains no segmental spectral information. The version of $s(t)$ used to train the WaveNet was sampled at 1000 Hz sample rate.

2.2. Network architecture

2.2.1. WaveNet implementation

A TensorFlow implementation of the WaveNet network architecture [1] was used in this work.

Briefly, the WaveNet is a feed-forward network that learns to generate conditional probabilities of quantized sample distributions given a sequence of the previous samples, using several stacked-up dilated convolutional layers. At generation time, the predicted sample (selected based on the predicted distribution) is directly fed back as part of the input of the network in an auto-regressive fashion. The stacked dilated convolutional layers increase the size of the receptive field for the prediction (i.e, the length of the previous portion of the signal that conditions sample generation). Gradually increasing dilation of each subsequent layer also provides a sort of parallel hierarchical analysis with more dilated layers capturing progressively longer-term dependencies in the signal.

In the present work we used a WaveNet network with two stacked-up sets, each containing 9 stacked layers with dilations incrementally doubled for each subsequent layer: 1, 2, 4, ..., 256, 1, 2, 4, ..., 256. This leads to the receptive field of length 1024 samples, corresponding to just over a second of the low sample rate signal. The network used 128 skip channels and 64 residual channels.

A μ-law companding transformation was applied to the prosodic signal to reduce the dynamic range. The network was trained to generate the processed signal quantized to 128 possible values.

In parallel to generating the prosodic signals, the network used here was also trained to produce a one-dimensional numerical "ramp" signal corresponding to individual syllables in the original speech signal. Each ramp linearly increases from 0 to 1 during the duration of the syllable, and is reset to 0 at the onset of the subsequent syllable. Training of this secondary target is implemented through separate post-processing layers and an output layer with regression loss added to the standard cross-entropy loss.

2.2.2. Embeddings

In addition to conditioning by the previous signal, the WaveNet architecture uses global conditioning to generate a signal with the required characteristics [1]. The known characteristics of a given signal (speaker's id, sentence type, ...) are fed as an additional input to each dilated convolutional layer via embeddings trained alongside the other network components. The embedding layer maps a discrete set of relevant parameters (individual characteristics of the signal or combinations thereof) to real-number valued vectors that are directly used to condition each convolutional layer.

The main aim of this paper is to evaluate whether these vector representations capture relevant characteristics of the signal elicited by external conditions (ambient noise level and type) and focus type. Two global conditioning embeddings (connected in a series) were incorporated in the WaveNet implementation used here (see also [3]).

The first embedding, referred to as *target embedding*, maps one-hot encoded category labels of interest to a 16-dimensional real-valued conditioning vector. The categories evaluated here were all possible combinations of ambient noise characteristics and sentence focus type. Three noise types at three noise levels each, plus "no noise" condition yield 10 noise characteristics combinations. When combined with 3 types of focus structure, the overall number of different conditioning categories is 30.

The creation of prosodic signals described above keeps intact some properties of the signal that might be interfering with the aims of our analysis, such as differences among speakers. To counteract this source of variability, we use a second embedding layer, called here *normalization embedding* conditioning the network through (16-dimensional) embedding of 21 speakers' IDs.

2.3. Training procedure

Models were trained for maximum of 100 epochs using Adam optimizer (learning rate of 0.001), with a single utterance per batch. Training set contained 75 % of the data, i.e., 3 randomly selected utterances from each condition of each speaker (out of 4 in the corpus, see Section 2.1.1).

Training was repeated 14 times, and for each trained model with separately randomized training data we extracted the embeddings.

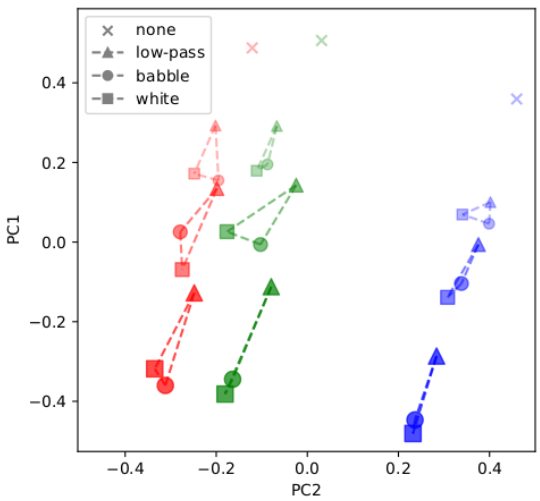

Figure 2: *An example of distribution of the embedding vectors of conditioning factor combinations, rendered in PC1-PC2 space. The noise level influence is indicated by color saturation and symbol size, from 0 dB to 80 dB (top to bottom). The colors indicate the focus condition influence, from left to right, N2 (red), broad focus (green) and N1 (blue).*

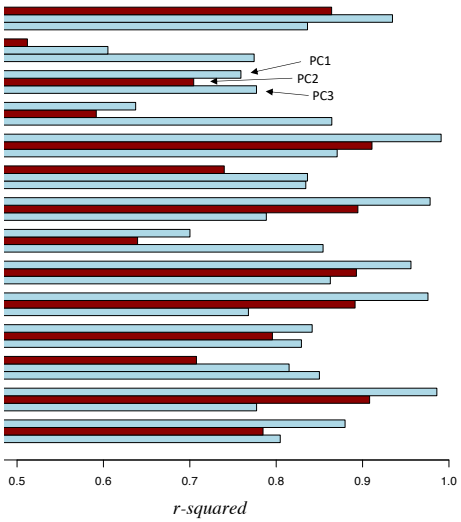

Figure 3: *The r-squared values (> 0.5) of the fits of noise level and focus condition against the first three PCs, for all 14 embeddings individually. The values for noise level are in dark red and for focus condition in light blue.*

3. Evaluation

3.1. Noise level and focus condition

The trained target embedding can be conceptualized as a matrix with the rows corresponding to different categories (in our case, 30 combinations of noise level, noise type and focus condition) and columns to 16 embedding dimensions. The embeddings for each trained WaveNet were first normalized by subtracting the mean value from each column of the embedding matrix.

The individual dimensions of the 16-dimensional embedding vectors cannot be assumed to contain readily interpretable information; different training instances can be expected to encode the relevant structural information contained in these vectors in a different way. Therefore, each trained embedding was transformed by Principal Component Analysis treating the dimensions as variables and embedded categories as observations. The principal components capturing the main sources of variance in each trained embedding were subsequently used for statistical analyses of the embeddings. Overall, the first principal component (PC1) explained on average around 30 %, the first two around (PC1 and PC2) 55 % and the first three components around 65 % of variance. The first 8 components accounted for approximately 90 % of variance on average.

Fig. 2 shows the embedded 30 categories (for one particular trained embedding) plotted in PC1–PC2 space. As can be seen in this case, the ambient noise conditions and focus type are reflected in the distribution of the condition embeddings, with PC1 primarily capturing the influence of noise level and PC2 separating the different focus types, with broad focus and N2 (narrow focus on the object) close together, and the N1 focus type further away. As shown by the dashed triangles, the effect of noise type is relatively similar for each focus type and noise level combination, with the embeddings of the low-pass filtered noise somewhat separated from the embeddings of the two other noise types.

In order to evaluate whether these observations hold for each trained embedding, we fitted linear models with noise level, noise type and focus, respectively, as independent factors and each individual principal component as a dependent variable, separately for each trained embedding, and used r-squared values of the fits as a measure of correspondence.

Fig. 3 summarizes the findings. For each trained embedding, the r-squared was relatively high (greater than 0.5) for the fit of one of the first three PCs against noise level, as well as for the fit of the remaining two of the first three PCs against focus type. The r-squared for all other combinations were generally negligible (less than 0.1). As seen in Fig. 3, which of the three PCs corresponded to noise type *versus* which PCs corresponded to focus type differed between the trained embeddings, but the pattern pertained for all instances.

No principal components showed any robust correspondence pattern with noise type (the r-squared of the fits was greater than 0.5 for only one embedding, for PC5).

3.2. Noise type

While the noise type does not systematically correspond to any particular embedding PCs, Fig. 2 suggests that the embeddings nevertheless capture noise type influence. For each individual noise level–focus type combination, the noise noise type influence patterns are relatively stable: the embedding vectors for babble and white noise are relatively close to each other with the vectors for the low-pass noise further apart.

In order to investigate this observation, we calculated mutual distances between the three noise type embeddings for each of these 9 noise level–focus combinations (excluding embeddings for 0 dB ambient noise). Euclidean distance in the coordinate space of the first three PCs of the embeddings was used as a distance measure.

Fig. 4 summarizes the obtained distance measures among the pairs of noise type embeddings for all ambient noise levels and each focus condition, pooled together for all trained embeddings. As can be seen, in all but one focus–level combination, the babble–white noise embedding distance is smaller than the distances between the embeddings for these noise types and the one for low-pass filtered noise. Also, the distances for low-pass–babble and low-pass–white combinations tend to increase

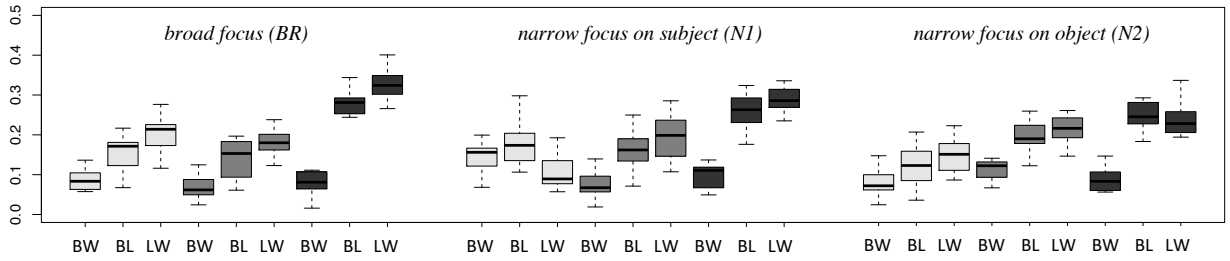

Figure 4: *Distances between embedding vectors for different noise types within the same noise level and focus conditions. The increasing noise level is represented by shade, from the lightest to the darkest grey capturing the levels from 60 to 80 dB. The Euclidean distances are calculated in PC1–3 space and pooled together for all embeddings. B = babble, W = white, L = low-pass filtered noise.*

with increasing ambient noise level, in particular for 80 dB noise, while the distance for babble–white noise pair remains relatively stable. These observations generalize the situation depicted in Fig. 2 for all trained embeddings.

The corresponding statistical linear model with embedding distance as a dependent variable, and noise type pair, noise level and focus type as independent factors, with all interactions, was fitted for the distance measurement data. The r-squared for the fit was 0.75.

The statistical model backs the observations listed above. With the exception of the 60 dB–N1 condition, the differences between the embedding distances for babble–white noise pair and the two other noise type combinations are significant ($p < 0.01$ for 60 dB in N2 focus condition, $p < 0.001$ for the rest). For the N1 and N2 focus types, the difference between babble–low-pass and low-pass–white types distances is not significant, but for broad focus condition it is ($p < 0.01$ for 60 dB; $p < 0.05$ for 70 and 80 dB levels).

The situation for N1 focus condition at 60 dB ambient noise is different. The distance between embeddings for low-pass and white noise type is significantly smaller than the distances for the two other combinations ($p < 0.001$ for comparison with babble–low-pass combination; $p < 0.05$ for comparison with the babble–white pair). The difference between the babble–white and babble–low-pass distances is not significant.

4. Discussion

As our aim was to present a new methodology, we primarily set out to replicate, and possibly extend the results of a previous study on the same material [4] rather than present radically new findings in the well-investigated area of Lombard speech.

As illustrated in Fig. 2, one of the principal components of the embedding space captures the effect of increasing noise level on speech signal in a way compatible with the results of [4], see also Fig. 1. The embedding space clearly shows the effect of focus type on signal, the effect that was in fact not captured in the previous work. Also, while our subsequent analysis of the influence of noise type largely agrees with the results reported in [4], it also suggests further interaction between the noise type effect and focus condition.

Statistical analysis of multiple instances of trained embeddings was used to identify properties of the embedding space that get repeatedly captured by the synthesis system. This approach allowed us to extract the influences that seem to be essential for the task of learning how to generate speech-like signals in the given conditions, at least by the given synthesis system. This technique deserves further exploration. What is the appropriate number of trained instances for the analysis? Can the final loss of the trained system be included in the statistical modelling? Will different synthesis systems yield compatible results? As we actually trained a generative synthesis system, can synthesised signals be used to provide new data for subsequent analysis?

In our opinion, the present approach provides several potential advantages. Although for this work we used some features extracted from the speech signal (f_0) and even annotations (syllable boundaries for the secondary target), in principle, these are not necessary. The system can be trained on the full signal, or a signal filtered and downsampled in a different way, without a secondary target. The present system can be thus used to analyze signals without the need for error-prone prosodic parameter extraction, with the errors potentially contaminating subsequent statistical analysis. Also, this approach does not require various standard statistical simplifications such as calculating averages over words, etc.

The embeddings presumably incorporate all information relevant for producing the generated signal, including, in our case frequency, intensity, temporal characteristics as well as mutual interactions among them. We can, however, restrict the signal characteristics to perform a more focused analyses. We have in fact trained the same synthesis system on several differently pre-processed signals, including full speech waveform, downsampled signals with constant f_0 and original envelopes, as well as on signals with original f_0 contours and stylised energy envelopes. We will report the results of comparative statistical analyses of the resulting embeddings in the near future.

Admittedly, there might be a limit on the types of hypotheses that can be tested in this way. One of the limiting factor is a relatively large size of the corpus required to train a WaveNet (or any other) statistical speech synthesis system. However, as shown in this work, meaningful analysis results can still be obtained with a considerably smaller corpus than would be required for a fully-fledged speech synthesis system.

Although not large, the material used in this work was well structured and tailor-made for analyzing Lombard speech. The requirement of an appropriate corpus structure (in terms of sufficient amount of samples with reasonably uniform realizations of the investigated effect) may be another constraining factor in terms of possible speech phenomena investigated in this way.

Finally, in our statistical analysis we used several standard statistical tools. It is possible that different statistical approaches might be more appropriate for quantitative evaluation of conditioning embedding spaces.

5. Acknowledgements

We would like to thank Vassilis Tsiaras from the University of Crete for providing us with the WaveNet implementation that was adapted for this work.

6. References

[1] A. v. d. Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, "Wavenet: A generative model for raw audio," *arXiv preprint arXiv:1609.03499*, 2016.

[2] T. Hayashi, A. Tamamori, K. Kobayashi, K. Takeda, and T. Toda, "An investigation of multi-speaker training for wavenet vocoder," in *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2017, pp. 712–718.

[3] A. Suni, M. Wlodarczak, M. Vainio, and J. Šimko, "Comparative analysis of prosodic characteristics using wavenet embeddings," in *Interspeech 2019, Graz, Austria, 15-19 September 2019*. The International Speech Communication Association (ISCA), 2019, pp. 2538–2542.

[4] M. Vainio, D. Aalto, A. Suni, A. Arnhold, T. Raitio, H. Seijo, J. Järvikivi, and P. Alku, "Effect of noise type and level on focus related fundamental frequency changes," in *Proc. Interspeech 2012*, Portland, Oregon, USA, 2012.

[5] E. Lombard, "Le signe de l'elevation de la voix," *Ann. Mal. de L'Oreille et du Larynx*, pp. 101–119, 1911.

[6] C. Rivers and M. Rastatter, "The effects of multitalker and masker noise on fundamental frequency variability during spontaneous speech for children and adults." *The Journal of auditory research*, vol. 25, no. 1, pp. 37–45, 1985.

[7] J. H. Hansen, "Analysis and compensation of stressed and noisy speech with application to robust automatic recognition," Ph.D. dissertation, Georgia Institute of Technology, 1988.

[8] R. Patel and K. W. Schell, "The influence of linguistic content on the Lombard effect," *Journal of Speech, Language, and Hearing Research*, vol. 51, pp. 209–220, 2008.

[9] J. H. Hansen and V. Varadarajan, "Analysis and compensation of lombard speech across noise type and levels with application to in-set/out-of-set speaker recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 2, pp. 366–378, 2009.

[10] A. Varga and H. J. Steeneken, "Assessment for automatic speech recognition: Ii. noisex-92: A database and an experiment to study the effect of additive noise on speech recognition systems," *Speech communication*, vol. 12, no. 3, pp. 247–251, 1993.