Jokinen, Emma; Alku, Paavo

Estimating the spectral tilt of the glottal source from telephone speech using a deep neural network

Published in: Journal of the Acoustical Society of America

DOI: 10.1121/1.4979162

Published: 01/04/2017

Please cite the original version:
Estimating the spectral tilt of the glottal source from telephone speech using a deep neural network

Emma Jokinen, and Paavo Alku

Citation: The Journal of the Acoustical Society of America 141, EL327 (2017); doi: 10.1121/1.4979162
View online: https://doi.org/10.1121/1.4979162
View Table of Contents: http://asa.scitation.org/toc/jas/141/4
Published by the Acoustical Society of America

Articles you may be interested in
- Streaming and sound localization with a preceding distractor
  The Journal of the Acoustical Society of America 141, EL331 (2017); 10.1121/1.4979167

- A finite element study on the cause of vocal fold vertical stiffness variation
  The Journal of the Acoustical Society of America 141, EL351 (2017); 10.1121/1.4978363

- The acoustic counterpart to coarticulation resistance and aggressiveness in locus equation metrics and vowel dispersion
  The Journal of the Acoustical Society of America 141, EL345 (2017); 10.1121/1.4979301

- Use of semantic context and $F_0$ contours by older listeners during Mandarin speech recognition in quiet and single-talker interference conditions
  The Journal of the Acoustical Society of America 141, EL338 (2017); 10.1121/1.4979565

- Experimental validation of acoustic intensity bandwidth extension by phase unwrapping
  The Journal of the Acoustical Society of America 141, EL357 (2017); 10.1121/1.4979604

- Relative contributions of formants to the intelligibility of sine-wave sentences in Mandarin Chinese
  The Journal of the Acoustical Society of America 141, EL495 (2017); 10.1121/1.4983747
Estimating the spectral tilt of the glottal source from telephone speech using a deep neural network

Emma Jokinen¹ and Paavo Alku
Department of Signal Processing and Acoustics, Aalto University, Espoo, Finland
emma.jokinen@aalto.fi, paavo.alku@aalto.fi

Abstract: Estimation of the spectral tilt of the glottal source has several applications in speech analysis and modification. However, direct estimation of the tilt from telephone speech is challenging due to vocal tract resonances and distortion caused by speech compression. In this study, a deep neural network is used for the tilt estimation from telephone speech by training the network with tilt estimates computed by glottal inverse filtering. An objective evaluation shows that the proposed technique gives more accurate estimates for the spectral tilt than previously used techniques that estimate the tilt directly from telephone speech without glottal inverse filtering.

1. Introduction
The overall spectral structure of voiced speech is low-pass but the slope of the spectrum, called the spectral tilt, varies due to factors such as the phonation type. Spectral tilt originates from the glottal excitation generated by vibrating vocal folds; smooth closing of the vocal folds results in a large spectral tilt and a more abrupt closing gives a glottal flow with a small spectral tilt. Spectral tilt has been analyzed, for instance, in studying linguistic stress, phonation type, vocal effort, and speech intelligibility improvement. While previous studies have measured the spectral tilt mainly in order to parameterize and classify speech production, tilt analysis can also be used to modify the speech signal, for instance, by changing its vocal effort characteristics. For this type of application, a parametric representation, such as an all-pole model, that can be used to directly modify the tilt is desirable. However, obtaining an all-pole model of the spectral tilt is not straightforward because direct, non-invasive acoustical recordings of the glottal source are not possible. Therefore, previously all-pole techniques based on estimating the spectral tilt directly from the speech signal, instead of the true glottal excitation, have been used. However, these methods might give tilt estimates that are distorted by poorly canceled formants.

The problem of spectral tilt estimation is akin to that of glottal inverse filtering (GIF), which aims to compute the time-domain waveform of the glottal flow from speech. Using GIF, the glottal flow is first estimated, after which its spectral tilt is parametrized with a low-order all-pole model. This approach gives tilt estimates that are free from vocal tract resonances because they have been removed by GIF. Unfortunately, GIF methods are vulnerable to degradations that are present in coded speech such as quantization noise and phase non-linearity. Therefore, GIF cannot be used directly in the estimation of spectral tilt from telephone speech. Nonetheless, applications—such as intelligibility enhancement in speech transmission—would benefit if the spectral tilt of the glottal source could be estimated from coded telephone speech using an all-pole model.

Therefore, a deep neural network (DNN) is proposed in the current study for the estimation of the spectral tilt of the glottal source from telephone speech. The DNN was trained using reference spectral tilt models computed from glottal flows estimated by a state-of-the-art GIF technique, quasi closed phase (QCP) GIF. The proposed method for estimating the spectral tilt of the glottal source with a DNN was compared in an objective evaluation to previously used methods that are based on estimating the tilt directly from the speech signal.

2. DNN-based spectral tilt estimation
To train the DNN, a speech database containing normal and Lombard speech from six Finnish speakers was used. From each of the six speakers in the database, 100
 utterances of both normal and Lombard speech were used to compute the features for the DNN. The input features were computed from simulated narrowband and wideband telephone speech, whereas the output features, modeling the reference spectral tilt, were computed from the original speech signals in the database that have been resampled at two sampling rates (8 and 16 kHz). These data are referred to as high-quality speech.

In order to model telephone speech, the data were pre-processed with the MSIN filter\(^7\) at 16 kHz for narrowband speech and with the HP50 filter\(^7\) at 48 kHz for wideband speech. Both filters are high-pass filters designed to simulate mobile station input characteristics. After the filtering, the speech samples were downsampled to either 8 or 16 kHz, and encoded and decoded using either the AMR narrowband codec\(^8\) at 12.2 kbps or the AMR wideband codec\(^9\) at 23.85 kbps. Finally, the samples were equalized to −26 dBov with SV56 (Refs. 7 and 10) after which the delay caused by the pre-processing was compensated for with an accuracy of one sample. The input features were computed using 20-ms frames with a 10-ms overlap from the telephone speech. The input features of the DNN were the logarithmic spectral magnitudes computed from the Hamming-windowed voiced frames using a 512-point fast Fourier transform (FFT). The dimensionality of the input features was 255, containing the first half of the FFT bins with the energy removed.

QCP was used in the computation of the DNN output features (describing the spectral tilt of the glottal source) from the high-quality speech. QCP calls for using an \(F_0\) estimator\(^11\) to determine the glottal closure instants (GCIs). After inverse filtering, the \(F_0\) computed from the GCI estimates was compared to the original \(F_0\) estimate and frames with large differences between the two values were discarded. To model the reference spectral tilt, the obtained glottal flow was differentiated and an all-pole filter was fitted to the Hamming-windowed signal. The all-pole filter was obtained using fourth order power-law adjusted linear prediction (LP-\(a\)),\(^12\) where the power spectrum is exponentiated to the power \(a = 1/2\) in this study. While the conventional linear prediction (LP) model of the differential glottal flow occasionally captures the so-called glottal formant, resulting in a sharp resonance in the spectrum, LP-\(a\) reduces this problem by compressing the spectrum. To match the original dynamic range of the spectrum, the final all-pole model is constructed by cascading two of the fourth order all-pole filters. The final output features for the training are the LP-\(a\) coefficients parametrized as line spectral frequencies (LSFs).

The number of features extracted in this manner was divided into train, test, and validation sets in such a way that 20% of the data were used for testing while the remaining 80% was further divided so that 80% of it was used for training the DNN and 20% for validation. The division of the data into training, validation, and test sets was randomized. The input features of the DNN were normalized to the range \([-1, 1]\) while the LSF output was in the range \([0, \pi]\). DNNs for both narrowband and wideband speech used standard feed-forward multilayer perceptrons with four hidden layers with sigmoid activations. The best results were obtained with layer sizes \([150, 100, 100, 50]\) for narrowband speech and \([100, 100, 50, 50]\) for wideband speech. The networks were randomly initialized and the training was done using standard backpropagation with early stopping to prevent overfitting. The mean squared test error was approximately 0.007 for narrowband speech and 0.008 for wideband speech.

3. Other spectral tilt estimation techniques

Three techniques—stabilized weighted linear prediction (SWLP),\(^13\) double linear prediction (2LP),\(^4\) and cepstral fit (CF)—were used as reference methods that all estimate the spectral tilt of the glottal source directly from speech.

SWLP is an all-pole modeling technique in which the square of the residual is temporally weighted based on the short-time energy of speech. Depending on the all-pole model order \((p)\) and the length of the temporal window \((M)\), relatively smooth models for the spectral tilt can be obtained. SWLP has been used previously for the mapping of the spectral tilt with Gaussian mixture models.\(^4\) The parameter values used in this study were \(M = 2\) and \(p = 6\).

The two-stage all-pole modeling method 2LP first computes a 20th-order LP inverse filter impulse response from the Hamming-windowed input frame and the obtained sequence is then used as an input in a sixth-order LP analysis. To remove any remaining effects of formants, the obtained impulse response is windowed with an exponential window. The 2LP method has been used to estimate and compensate for the spectral tilt of voiced speech.\(^4\)
In the CF technique, a 12th-order all-pole model is first computed from the Hamming-windowed input frame and the corresponding spectrum is used to obtain the real cepstrum. Next, the first two coefficients of the cepstrum are transformed into an autocorrelation sequence from which a sixth-order all-pole model is computed.

4. Evaluation

To evaluate the performance of the proposed spectral tilt estimation technique, an objective evaluation was conducted using a data set of Finnish vowels produced in different phonation types (breathy, normal, and pressed). The data set contains samples from 12 speakers (six male). In the evaluation, QCP was used to estimate the glottal flow by computing the required \( F_0 \) value from the accompanying electroglottography signal. The glottal flow was differentiated and used to estimate the reference spectral tilt using the LP-\( z \) method. For the evaluation, the vowel samples were pre-processed to model telephone speech. After this, the spectral tilt was estimated in 20-ms frames with 50\% overlap using either the proposed DNN-based technique or one of the reference techniques (SWLP, 2LP, or CF). The accuracy of the spectral tilt estimation was measured using the logarithmic spectral distortion (logSD) of the tilt estimates.

Examples of the spectral tilts estimated using different techniques for both narrowband and wideband frames are shown in Fig. 1. The results of the objective measures for both narrowband and wideband speech are shown in Table 1. The results indicate that while the other techniques, especially the LP-based ones (SWLP and 2LP), tend to follow the resonances of the vocal tract, thus leading to high logSD scores, the proposed DNN-based technique is able to capture the real glottal source spectral tilt more closely. The objective results, which are comparable to the 6-dB logSD errors observed in artificial bandwidth extension, confirm that in most cases the DNN-based method performs better in estimating the spectral tilt originating from the glottal flow.

![Fig. 1. An example demonstrating different spectral tilt estimation methods for both narrowband and wideband speech. The upper panels show the spectra computed from the coded speech (Speech), the glottal flow derivative estimated from the high-quality speech \( \frac{d}{dt} \text{flow} \) and the reference all-pole spectral tilt model (Ref.). The lower panels show the spectra of the different tilt estimation techniques (DNN, SWLP, 2LP, and CF).](image)

Table 1. Comparison of the spectral tilt estimation techniques using objective metrics for both narrowband and wideband speech. The techniques under comparison are the proposed DNN-based method (DNN), 2LP, SWLP, and CF. The objective measures are the logSD (in dB) of the spectral tilt estimates between the methods and the reference obtained using GIF. The spectral tilt estimates are evaluated for breathy (B), normal (N), and pressed (P) phonation types. The lowest error for each column has been highlighted in bold font.

<table>
<thead>
<tr>
<th></th>
<th>Narrowband</th>
<th></th>
<th>Wideband</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>B</td>
<td>N</td>
<td>P</td>
<td>B</td>
</tr>
<tr>
<td>DNN</td>
<td>8.02</td>
<td>6.25</td>
<td>5.82</td>
<td>8.39</td>
</tr>
<tr>
<td>2LP</td>
<td>18.35</td>
<td>12.36</td>
<td>8.81</td>
<td>26.49</td>
</tr>
<tr>
<td>SWLP</td>
<td>11.36</td>
<td>8.16</td>
<td>7.09</td>
<td>11.26</td>
</tr>
<tr>
<td>CF</td>
<td>13.36</td>
<td>9.38</td>
<td>7.73</td>
<td>18.19</td>
</tr>
</tbody>
</table>

In the CF technique, a 12th-order all-pole model is first computed from the Hamming-windowed input frame and the corresponding spectrum is used to obtain the real cepstrum. Next, the first two coefficients of the cepstrum are transformed into an autocorrelation sequence from which a sixth-order all-pole model is computed.
5. Conclusions

The spectral tilt of speech originates from the glottal flow generated by vibrating vocal folds. The glottal flow cannot be measured directly from human speech production but it can be estimated using an inversion methodology, GIF. Unfortunately, GIF analysis calls for having high-quality speech as input. Therefore, the GIF-based estimation of the spectral tilt of the glottal source has not been possible previously from coded telephone speech. In order to tackle this shortcoming, a DNN-based technique is proposed for the estimation of the spectral tilt of the glottal source from telephone speech. The DNN, whose input features are computed from coded telephone speech, is trained using spectral tilt estimates computed with a state-of-the-art GIF technique using high-quality speech. Using an objective metric, the DNN-based method is compared with three known LP-based spectral tilt estimation techniques that compute estimates directly from speech without using GIF. The results indicate that the proposed method produces smaller errors in all of the evaluated conditions compared to the LP-based techniques. The application with the most potential for the proposed technique is speech transmission where the method could be used in mapping the original spectral tilt of the received coded signal onto a new, smaller tilt, hence improving speech intelligibility in noisy near-end conditions.4

Acknowledgments

This research has been funded by the Academy of Finland (Project No. 284671).

References and links