Modeling Formant Dynamics in Speech Spectral Envelopes

Alexandra Craciun*, Jouni Paulus†, Gökhan Sevkin‡ and Tom Bäckström§
*International Audio Laboratories Erlangen, Friedrich-Alexander University (FAU), Erlangen, Germany,
currently with XMOS Ltd, Bristol, UK
†Fraunhofer Institute for Integrated Circuits (IIS), Erlangen, Germany
‡International Audio Laboratories Erlangen, Friedrich-Alexander University (FAU), Erlangen, Germany
§Aalto University, Department of Signal Processing and Acoustics, Espoo, Finland

Abstract—**The spectral envelope of a speech signal encodes information about the characteristics of the speech source. As a result, spectral envelope modeling is a central task in speech applications, where tracking temporal transitions in diphones and triphones is essential for efficient speech synthesis and recognition algorithms. Temporal changes in the envelope structure are often derived from estimated formant tracks, an approach which is sensitive to estimation errors. In this paper we propose a speech source model which estimates frequency and amplitude movements in the spectral envelopes of speech signals and does not rely on formant tracking. The proposed model estimates the amplitude and frequency shifts for each sub-band and time frame of a speech signal using the information from the previous time frame. Our experiments demonstrate that the model captures temporal structures of spectral envelopes with high precision. The proposed model can thus be applied as an accurate low-order representation of temporal dynamics in speech spectral envelopes.**

I. INTRODUCTION

Continuity is a fundamental property of speech, which is supported for instance by the fact that there are no specific acoustic markers or gaps for delimiting phonemes [1]. Switching from one phoneme to another is usually a smooth transition, where the movement of a phoneme also affects the moving pattern of the following phoneme [2], an effect better known as *coarticulation*. This suggests that features which describe the smooth transitions of speech over time carry important additional information compared to *static* speech features.

In this paper we propose a speech source model which captures smooth temporal movements of speech spectral envelopes in both amplitude and frequency. The estimated spectral shifts can thus be viewed as *dynamic* formant movements. While typical formant movement calculation relies on LPC (Linear Predictive Coding)-based formant tracking, our approach does not require formant tracking. Thus, it is unaffected by typical formant tracking problems such as spectral peak estimation, where it can happen that the wrong peak position is selected due to small envelope variations around the true peak. Also, in contrast to conventional formant tracking approaches, where only the position of the spectral peaks is estimated, our model takes into account the movement of the entire spectral envelope shape in both frequency and amplitude. This allows tracking envelope movements also in areas with not so prominent peaks (where the peaks are very smooth). Furthermore, since it is not limited to the few most prominent peaks, but rather analyzes the entire area around the peaks, it is able to track the spectral movement around peaks which are close together. In the following, we will discuss the applications for formant tracking and dynamic formant movement estimation.

Formant tracking is an indispensable tool in speech applications, in particular in automatic speech recognition, because it is easy to identify important groups of phonemes, e.g., vowels, using only the formant information [3]. Other applications such as text-to-speech synthesis [4] or spectral amplification of speech in hearings aids [5] also rely on formant estimation. Although robust for identifying vowels or certain consonants such as nasal consonants, formants cannot be used to discriminate between the rest of the consonants. Another drawback of formant tracking algorithms is the fact that they require peak picking for locating the formants. Peak picking algorithms are sensitive to small local variations in the envelope contour, which can lead to picking a local maximum instead of the global one. In the proposed method, we avoid peak picking altogether by using a minimum sum of squared residuals to track the movements of the entire spectral envelope.

Dynamic formant movements are important acoustic cues for speech analysis and synthesis. For instance, in speech synthesis, including formant dynamics and using the original formant contours instead of flat ones (constructed from the formant frequencies measured in the steadiest part of a vowel) was shown to improve vowel intelligibility [6]. In speech analysis, large formant movements were observed in hyperarticulated speech, i.e., speech emphasized by a talker in a noisy environment or when speaking to a hearing-impaired person [7]. Since such dynamics were not observed for normal or conversational speech, these differences allow distinguishing between the two types of speech. Dynamic formant movements are also employed in vowel identification tasks, for which metrics based on formant movements, steady-state formant values and vowel duration are used [8]. Due to the design based on LPC formant tracking, the previous formant movement metrics suffer from several drawbacks. For instance, spectral change estimators cannot efficiently describe more complex formant movements, while the spectral angle

ISBN 978-0-9928626-7-1 © EURASIP 2017

cannot capture the direction of the formant movement [8]. In our proposed approach, we avoid such issues since we analyze the amplitude and frequency movements over the entire spectral envelope in time and do not restrict ourselves to a few frequency points for the main formants. An additional advantage of the proposed formant shift estimation is the fact that we can separately investigate the movement in frequency and the movement in amplitude.

Another application for both formant tracking and dynamic formant movement is in medical speech analysis, where they can be used for detecting certain motor neurone diseases, such as amyotrophic lateral sclerosis (ALS), which is characterized at an early age by speaking difficulties. Identifying speech properties that correlate with such abnormal speaking behavior is crucial for a quick and early detection of the disease. In [9], the authors propose using features based on formant statistics and on the first and second derivatives of formant trajectories. They show that the mean second formant (F2) speed and the mean F2 acceleration have the highest correlation with the intelligibility and speaking rate of ALS patients. Our formant shift estimate can be flexibly modified to include only the subbands corresponding to the F2 range and can thus be directly used for such detecting ALS without the need of performing formant tracking.

Such a speech source model is also relevant for all applications which require low-order representations of the spectral envelope. For instance, the information contained in the estimated formant shift parameters can be used to interpolate lost speech frames in case of packet loss for voice communication applications [10].

II. Signal Model

The proposed source model is not a global model, but rather a local one, since it analyzes spectral envelope movements over a small frequency range, where we assume that the temporal change in the envelope is constant. This assumption is necessary since the formant shift may vary from formant to formant. That is, over time Δt, the first formant might move in the opposite direction compared to the second formant. However, if we consider a small enough frequency region, e.g., over only one formant, the formant shift is expected to be constant (all points move by the same distance in the same direction). A safe choice for the frequency range is the bandwidth of the narrowest formant. Let us consider such a small frequency range as depicted in Fig. 1, where $\mathbf{k_i}$ represents the vector of all frequency bins over which the formant shapes stretch, while k_0 and $k_0 + \Delta k$ denote each a frequency bin index. The horizontal axis in the figure shows the frequency, while the vertical axis shows the spectral envelope amplitude. Note that vectors are marked by bold letters, while normal letters denote scalars.

In the proposed model, the constant formant shift, which is depicted by a diagonal direction in Fig. 1, is decomposed into two movements: a vertical one, denoted by $a_i\Delta t$, and a horizontal one, denoted by $b_i\Delta t$, where a_i and b_i are real-valued scalars. Our aim is to model the formant shift for the time transition $t_0 \rightarrow t_0 + \Delta t$ as a function of the current and previous envelopes. Let us consider the example shown in Fig. 1, where the points on the curve $\mathbf{X}(\mathbf{k_i}, t_0)$ move by the same distance (depicted by diagonal arrows) to new positions on the curve $\mathbf{X}(\mathbf{k_i} + \Delta k, t_0 + \Delta t)$ at $t_0 + \Delta t$. $\mathbf{X}(\mathbf{k}, t)$ is the general expression for the magnitude spectral envelope of a speech signal and $\mathbf{k}$ is the vector of frequency bins up to the Nyquist frequency. Thus, if the peak maximum $X(k_0, t_0)$ denotes a formant, at time $t_0 + \Delta t$ the formant shifts to $X(k_0 + \Delta k, t_0 + \Delta t)$, where $\Delta k = b\Delta t$. For notational simplicity, we will neglect the index i in the following.

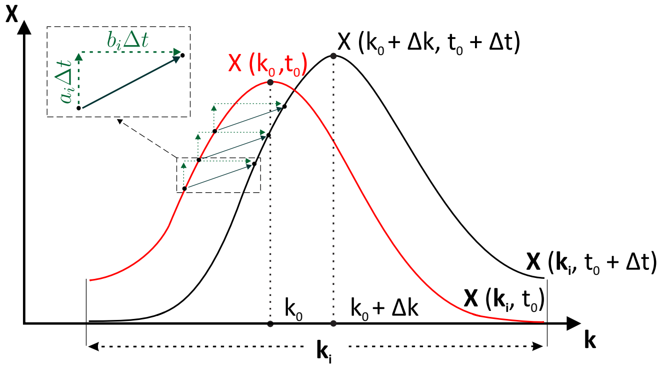

Fig. 1. Spectral envelope shift $\mathbf{X}(\mathbf{k_i}, t_0) \rightarrow \mathbf{X}(\mathbf{k_i}, t_0 + \Delta t)$ is assumed constant over small frequency range $\mathbf{k_i}$.

Let us focus on the formant shift from $X(k_0, t_0)$ to $X(k_0 + \Delta k, t_0 + \Delta t)$ and write $X(k_0 + \Delta k, t_0 + \Delta t)$ as a function of $X(k_0, t_0)$:

$$\begin{aligned} X(k_0 + b\Delta t, t_0 + \Delta t) &= (1 + a\Delta t) \cdot X(k_0 + \Delta k, t_0) \\ &\simeq (1 + a\Delta t) \cdot [X(k_0, t_0) + \Delta k X'(k_0, t_0)] \end{aligned} \tag{1}$$

In Eq. 1, $X(k_0 + \Delta k, t_0)$ was approximated by a first-order Taylor series using $X(k_0 + \Delta k, t_0) \simeq X(k_0, t_0) + \frac{X'(k_0,t_0)}{1!}\Delta k$, where $X'(k_0, t_0)$ represents the derivative of $X(k_0, t_0)$ over frequency. The derivative of the magnitude spectral envelope can be computed by taking the gradient of the envelope over frequency or can be approximated by the simple difference between neighbouring points on the envelope.

For simplicity, we use X_0 instead of $X(k_0, t_0)$ and $\widehat{X}_1$ instead of $X(k_0 + \Delta k, t_0 + \Delta t)$ and replace Δk by $b\Delta t$ in Eq. 1, which, considering the entire frequency range $\mathbf{k_i}$, becomes:

$$\widehat{\mathbf{X}}_1 \simeq (1 + a\Delta t) \cdot (\mathbf{X}_0 + b\Delta t \cdot \mathbf{X}'_0). \tag{2}$$

Using Eq. 2, the formant shift parameters a and b are estimated by minimizing the sum of squared residuals $E_{\text{res}}(\widehat{\mathbf{X}}_1, \mathbf{k_i})$, where the residual is computed as the difference between the estimated and the true envelope at $t_0 + \Delta t$ over the frequency range denoted by $\mathbf{k_i}$:

$$\begin{aligned} \min_{a,b} E_{\text{res}}(\widehat{\mathbf{X}}_1, \mathbf{k_i}) &= \min_{a,b} ||\mathbf{X}_1 - \widehat{\mathbf{X}}_1||^2 \\ &= \min_{a,b} ||\mathbf{X}_1 - \left[(1 + a\Delta t) \cdot (\mathbf{X}_0 + b\Delta t \cdot \mathbf{X}'_0)\right]||^2. \end{aligned} \tag{3}$$

Here $\widehat{\mathbf{X}}_1$ corresponds to our model approximation of $\mathbf{X}_1$ in Eq. 2 and $||\mathbf{X}||^2 = \mathbf{X}^T\mathbf{X}$ denotes the vectorial 2-norm.

The previous equation is further simplified by the variable substitution $\mathrm{c} = (1 + \mathrm{a}\Delta\mathrm{t})$ and $\mathrm{d} = \mathrm{b}\Delta\mathrm{t} \cdot (1 + \mathrm{a}\Delta\mathrm{t})$, resulting in:

$$\min_{\mathrm{a,b}} ||\mathbf{X}_1 - \widehat{\mathbf{X}}_1||^2 = \min_{\mathrm{c,d}} ||\mathbf{X}_1 - \mathrm{c}\mathbf{X}_0 - \mathrm{d}\mathbf{X}_0'||^2. \tag{4}$$

This can be rewritten in a simpler form as:

$$\min_{\mathrm{a,b}} ||\mathbf{X}_1 - \widehat{\mathbf{X}}_1||^2 = \min_{\mathbf{s}} ||\mathbf{X}_1 - \mathbf{Z}\mathbf{s}||^2, \tag{5}$$

where $\mathbf{Z} = [\mathbf{X}_0 \quad \mathbf{X}_0']$ and $\mathbf{s} = [\mathrm{c} \quad \mathrm{d}]^T$. Eq. 5 can now be solved using the linear least squares approach, which results in $\mathbf{s} = (\mathbf{Z}^T\mathbf{Z})^{-1}\mathbf{Z}^T\mathbf{X}_1$, which can be expanded into:

$$\begin{bmatrix} \mathbf{s} \end{bmatrix} = \begin{bmatrix} \mathrm{c} \\ \mathrm{d} \end{bmatrix} = \begin{bmatrix} \mathbf{X}_0^T\mathbf{X}_0 & \mathbf{X}_0^T\mathbf{X}_0' \\ \mathbf{X}_0'^T\mathbf{X}_0 & \mathbf{X}_0'^T\mathbf{X}_0' \end{bmatrix}^{-1} \cdot \begin{bmatrix} \mathbf{X}_0^T\mathbf{X}_1 \\ \mathbf{X}_0'^T\mathbf{X}_1 \end{bmatrix}. \tag{6}$$

Using the definition of c and d, the final solution for a and b becomes:

$$\begin{cases} \mathrm{a} = \dfrac{1}{\Delta\mathrm{t}} \cdot \left(\dfrac{\mathbf{X}_0'^T\mathbf{X}_0'\mathbf{X}_0^T\mathbf{X}_1 - \mathbf{X}_0^T\mathbf{X}_0'\mathbf{X}_0'^T\mathbf{X}_1}{\mathbf{X}_0^T\mathbf{X}_0\mathbf{X}_0'^T\mathbf{X}_0' - \mathbf{X}_0^T\mathbf{X}_0'\mathbf{X}_0'^T\mathbf{X}_0} - 1 \right) \\ \mathrm{b} = \dfrac{1}{\Delta\mathrm{t}} \cdot \dfrac{-\mathbf{X}_0'^T\mathbf{X}_0\mathbf{X}_0^T\mathbf{X}_1 + \mathbf{X}_0^T\mathbf{X}_0\mathbf{X}_0'^T\mathbf{X}_1}{\mathbf{X}_0'^T\mathbf{X}_0'\mathbf{X}_0^T\mathbf{X}_1 - \mathbf{X}_0^T\mathbf{X}_0'\mathbf{X}_0'^T\mathbf{X}_1} \end{cases}. \tag{7}$$

III. System Description

The input time signal $\mathbf{x}(\mathrm{t})$ is first cut into 128 ms segments $\mathbf{x}_i$ with 75% overlap, where index i denotes the segment index. Each segment is multiplied with a Hamming window $\mathbf{w}$, resulting in $\mathbf{x}_{w,i}$, followed by the extraction of 16 LPC coefficients $\mathbf{x}_{LPC,i}$. A 1024-point Fourier transform is then applied over the LPC coefficients and the spectral envelope $\mathbf{X}_i$ is computed as the energy-normalized inverse magnitude thereof:

$$\begin{aligned} \mathbf{x}_{LPC,i} &= LPC\{\mathbf{x}_{w,i}\} \\ \mathbf{X}_{LPC,i} &= 1/\mathcal{F}\{\mathbf{x}_{LPC,i}\} \\ \mathbf{X}_i &= \mathbf{X}_{LPC,i} \cdot ||\mathbf{x}_{w,i}||^2/||\mathbf{X}_{LPC,i}||^2. \end{aligned} \tag{8}$$

The spectral envelope is then split into uniform 250 Hz sub-bands with 50% overlap. The sub-band overlap allows for a smooth combination of the estimates and avoids discontinuities around the boundaries of the sub-bands. For each sub-band, the envelope parameters a and b are estimated using Eq. 7. The estimated spectral envelope is then computed using Eq. 2. Note that due to the overlap between sub-bands, we need to combine several envelope estimates for the same frequency range. To do so, we multiply each sub-band by a Hanning window equal in length to the number of frequency bins in the sub-band and use an approach similar to the overlap-add method for the Short-Time Fourier Transform (STFT) to synthesize the envelope estimates into one single curve.

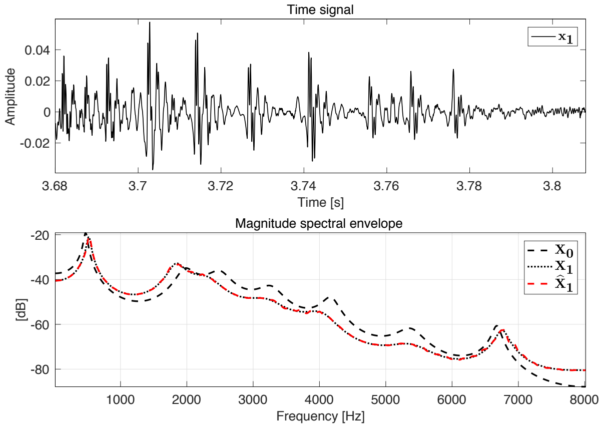

Fig. 2. Vowel-vowel transition. The time segment corresponding to the current frame is shown in the upper plot. The magnitude spectral envelopes of the previous frame (dashed black line) and current frame (dotted black line), together with the estimated spectral envelope of the current frame (dashed red line) are shown in the lower plot. Notice that shape of the estimated spectral envelope of the current frame follows so closely the shape of the true envelope of the current frame, that they are almost indistinguishable in the plot.

IV. Evaluation

In this section we will present the results of different evaluations on the proposed speech model. More precisely, our aim was to investigate the following:

- whether the model can accurately be used to reconstruct a speech envelope,
- a good choice for the sub-band bandwidth over which the model parameters are estimated such that both high accuracy and low complexity are achieved,
- whether there is a difference between modeling female and male speakers with the proposed speech model.

A. *Modeling accuracy*

In Fig. 2, a vowel-vowel transition frame is shown for item SA1 from \TEST\DR1\of the TIMIT database [11]. The file consists of a 16-bit, 16 kHz speech utterance from a female speaker in American English. In the upper plot, the current time frame is plotted. In the lower plot, the spectral envelope $\mathbf{X}_0$ of the previous frame (dashed black line) and the spectral envelope $\mathbf{X}_1$ of the current frame (dotted black line) are shown. The reconstructed envelope $\widehat{\mathbf{X}}_1$ of the current frame, using the estimated amplitude and frequency shift parameters a and b as described in Eq. 2, is plotted here as a dashed red line. Both amplitude and frequency shifts are clearly visible, e.g., looking at the first peak (formant) at 470 Hz, which in the current frame was shifted to 531 Hz. We note that the reconstructed spectral envelope follows very closely the true spectral envelope of the current frame. This suggests that the proposed source model is able to accurately incorporate the time evolution of the spectral envelope shape between the previous and the current frame.

To better understand how well the reconstructed spectral envelope matches the true envelope shape, we computed the

ISBN 978-0-9928626-7-1 © EURASIP 2017

estimation error $\mathbf{E_{res}}(\widehat{\mathbf{X}}_1, \mathbf{k})$, as described in Eq. 3, over the entire frequency range and for each frame. However, the absolute error is not a good indicator of how close our estimation is with respect to the true value. We therefore computed the relative error $\mathbf{E_{res,rel}}(\widehat{\mathbf{X}}_1, \mathbf{k}) = ||\mathbf{X}_1 - \widehat{\mathbf{X}}_1||^2/||\mathbf{X}_1||^2$, which contains the normalization with respect to $\mathbf{X}_1$. By replacing the estimated envelope of the current frame with the true envelope of the previous frame, we also computed the baseline relative error $\mathbf{E_{res,rel}}(\mathbf{X}_0, \mathbf{k}) = ||\mathbf{X}_1 - \mathbf{X}_0||^2/||\mathbf{X}_1||^2$. The baseline relative error serves as lower reference, describing the case where no estimate is available and the envelope of the previous frame is used instead. To our knowledge, there are no other methods which would be suitable to use for direct comparison with the estimated formant shift parameters. The delta-formant tracks or the time differences between the positions of the main formant tracks, which can be computed as in [12] is limited only to the position of the formants and cannot be used to model the areas in between formants for instance. In contrast, our approach is capable of modeling the movements along the entire spectral envelope. From this perspective, it might seem more plausible to compare our approach to a delta-envelope measure, which captures the difference in the spectral envelope over consecutive frames. Nevertheless, this does not allow separating between the two effects, i.e., the amplitude and the frequency shift, which is done in our model. As a result, a direct comparison of the proposed formant shift estimates with other features was not possible, so we chose to evaluate the envelope reconstruction error instead. Here, a plausible lower reference was easily identified by employing the envelope of the previous frame, which covers the scenario where no estimate is available and the closest match to the current frame is considered.

Fig. 3 shows in the upper plot the magnitude spectral envelope of item SA1. In the middle plot, the relative error of the envelope estimate (solid black line) and the relative baseline error (dashed black line) are shown. The relative error for our proposed model is on average -25.6 [dB], which is significantly lower than the average baseline error of -5.9 [dB]. In the lower plot, the difference between the two errors $\Delta\mathbf{E_{res,rel}} = \mathbf{E_{res,rel}}(\widehat{\mathbf{X}}_1, \mathbf{k}) - \mathbf{E_{res,rel}}(\mathbf{X}_0, \mathbf{k})$ is depicted. We can clearly see that the relative error for the estimated envelope is on average around 20 dB lower than the relative baseline error. This shows that our modeling is accurate and results in low residual error.

B. *Choice of sub-band bandwidth*

A general note on the proposed speech source model is that it is sensitive to the approach used for splitting the spectral envelope into sub-bands. The formant shift parameter estimation is done after the spectral envelope is extracted, so it is performed on a smooth curve, which is divided into smaller equal segments by means of a filterbank, where the amount of overlap and the bandwidth of each filter is flexible. The amplitude and frequency shifts of the formants are then estimated for each filter in the filterbank or sub-band. On one hand, the amount of overlap between the sub-bands is important since larger overlap allows recombining the sub-band estimates into a smooth spectral envelope. On the other hand, the size of each sub-band is critical for the proposed model since we need to make sure that the range of a sub-band is small enough such that our model assumption of having a constant spectral shift still holds. From informal tests, we observed that increasing the amount of sub-band overlap results in only small improvements in $\Delta\mathbf{E_{res,rel}}$. However, varying the sub-band bandwidth (BW) has a larger impact, which we shall discuss in more detail in the following.

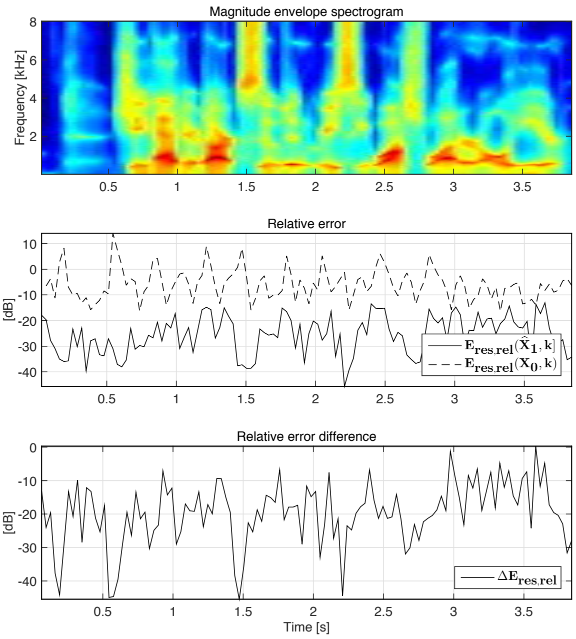

Fig. 3. The upper plot shows the magnitude spectral envelope of the audio signal. The middle plot shows the relative error between the envelope reconstructed with the model parameters and the true envelope (solid black line). The relative baseline error between the current and the previous spectral envelope is depicted here as a dashed black line. The bottom plot shows the relative error difference. The estimation was done for a sub-band bandwidth of 250 Hz.

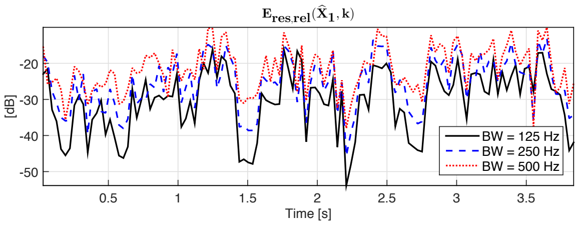

Fig. 4. The relative error between the envelope reconstructed with the model parameters and the true envelope for sub-band bandwidths of 125 Hz, 250 Hz and 500 Hz.

In Fig. 4, the error $\mathbf{E_{res,rel}}(\widehat{\mathbf{X}}_1, \mathbf{k})$ is depicted for sub-band bandwidths of 125 Hz (solid black line), 250 Hz (dashed blue line) and 500 Hz (dotted red line). We notice that the error gets smaller for narrower bandwidths, as expected.

The average error $\overline{\mathrm{E}_{\mathrm{res,rel}}}_{\mathrm{ALL,TIMIT}}$ for the proposed model and for the baseline over all 6299 items in TIMIT is shown

BW [Hz]	75	125	250	500	1000
$\overline{\mathrm{E}_{\mathrm{res,rel}}}_{\ \mathrm{ALL,TIMIT}}(\widehat{\mathbf{X}}_1, \mathbf{k})$ [dB]	-37.3	-29.6	-24.6	-21.2	-19.0
$\overline{\mathrm{E}_{\mathrm{res,rel}}}_{\ \mathrm{ALL,TIMIT}}(\mathbf{X}_0, \mathbf{k})$ [dB]	-5.4				

TABLE I
AVERAGE RELATIVE ERROR OF THE PROPOSED METHOD AND OF THE BASELINE OVER ALL TIMIT ITEMS FOR SUB-BAND BANDWIDTHS OF 75, 125, 250, 500 AND 1000 HZ.

BW [Hz]	75	125	250	500	1000
$\overline{\mathrm{E}_{\mathrm{res,rel}}}_{\ \mathrm{F,TIMIT}}(\widehat{\mathbf{X}}_1, \mathbf{k})$ [dB]	-39.4	-31.2	-25.9	-22.2	-19.9
$\overline{\mathrm{E}_{\mathrm{res,rel}}}_{\ \mathrm{F,TIMIT}}(\mathbf{X}_0, \mathbf{k})$ [dB]	-5.6				

TABLE II
AVERAGE RELATIVE ERROR OF THE PROPOSED METHOD AND OF THE BASELINE OVER ALL FEMALE SPEAKER TIMIT ITEMS FOR SUB-BAND BANDWIDTHS OF 75, 125, 250, 500 AND 1000 HZ.

BW [Hz]	75	125	250	500	1000
$\overline{\mathrm{E}_{\mathrm{res,rel}}}_{\ \mathrm{M,TIMIT}}(\widehat{\mathbf{X}}_1, \mathbf{k})$ [dB]	-36.3	-28.8	-24.0	-20.8	-18.6
$\overline{\mathrm{E}_{\mathrm{res,rel}}}_{\ \mathrm{M,TIMIT}}(\mathbf{X}_0, \mathbf{k})$ [dB]	-5.2				

TABLE III
AVERAGE RELATIVE ERROR OF THE PROPOSED METHOD AND OF THE BASELINE OVER ALL MALE SPEAKER TIMIT ITEMS FOR SUB-BAND BANDWIDTHS OF 75, 125, 250, 500 AND 1000 HZ.

in Tab. I for a wider range of sub-band bandwidths. For each item in TIMIT, the mean $\mathbf{E}_{\mathbf{res,rel}}$ over all frames in the item is computed and the average thereof over all items denotes the average relative error $\overline{\mathrm{E}_{\mathrm{res,rel}}}_{\mathrm{ALL,TIMIT}}$. The smallest error is obtained for the shortest sub-band (75 Hz). In comparison with the previously discussed case of 250 Hz, the improvement for a bandwidth of 75 Hz is 12.7 dB. While decreasing the sub-band bandwidth clearly results in better spectral shift estimates and thus a more accurate envelope estimate, this comes at the expense of a higher computational cost, since a larger number of parameters needs to be estimated. For instance, choosing 75 Hz instead of 250 Hz requires computing four times as many sub-bands, which means four times more estimation operations are performed. We therefore consider that the 250 Hz bandwidth is a good balance between estimation accuracy and estimation cost.

C. Modeling of female and male speakers

The proposed source model was also assessed separately for male and female speakers since formant positions vary for different genders. The average relative error for female and male speakers in TIMIT is depicted in Tab. II and Tab. III, respectively. The estimation error is between 1 dB and 3 dB lower for female speakers compared to male speakers, with larger differences observed for narrower sub-bands, e.g., 75 Hz. Formant frequencies of female speakers are on average higher than for male speakers [13], so the spectral envelopes of female voices should be smoother at lower frequencies. Since tracking changes in smooth envelopes is less susceptible for estimation errors, a slightly smaller estimation error of the formant dynamics is expected for female speakers.

V. CONCLUSIONS

In this paper we proposed a speech source model for estimating frequency and amplitude movements of the spectral envelope over time. In our investigations, we showed that the proposed source model can efficiently capture the temporal movements in the spectral envelopes. The high model accuracy was confirmed by a low relative error between the reconstructed envelope using the estimated parameters and the true envelope. For female speakers the average relative error was even a few dB lower than for male speakers. The performance of the model depends however on the sub-band bandwidth, which controls the balance between estimation accuracy and computational cost. The proposed speech source model is relevant for a variety of applications, in particular for those which require low-order representations of the spectral envelope variation over time.

REFERENCES

[1] P. Taylor, *Text-to-Speech Synthesis*. Cambridge University Press, 2009.

[2] J. Benesty, M. M. Sondhi, and Y. Huang, *Springer Handbook of Speech Processing*. Springer-Verlag Berlin Heidelberg, 2008.

[3] J. Franco-Pedroso and J. Gonzalez-Rodriguez, "Linguistically-constrained formant-based i-vectors for automatic speaker recognition," *Speech Communication*, vol. 76, 2016.

[4] W. Ding and N. Campbell, "Optimizing unit selection with voice source and formants in the CHATR speech synthesis system," in *Proceedings of the 5th European Conference on Speech Communication and Technology (EUROSPEECH '97)*, Rhodes, Greece, 1997.

[5] J. Schilling, R. Miller, M. Sachs, and E. Young, "Frequency-shaped amplification changes the neural representation of speech with noise-induced hearing loss," *Hearing Research*, vol. 117, no. 12, 1998.

[6] J. Hillenbrand and T. M. Nearey, "Identification of resynthesized /hVd/ utterances: Effects of formant contour," *The Journal of the Acoustical Society of America*, vol. 105, 1999.

[7] S. Ferguson and D. Kewley-Port, "Vowel intelligibility in clear and conversational speech for normal hearing and hearing-impaired listeners," *The Journal of the Acoustical Society of America*, vol. 112, 2002.

[8] S. Ferguson and D. Kewley-Port, "Talker differences in clear and conversational speech: acoustic characteristics of vowels," *Journal of Speech, Language and Hearing Research*, vol. 50, 2007.

[9] R. Horwitz-Martin, T. Quatieri, A. Lammert, J. Williamson, Y. Yunusova, E. Godoy, D. Mehta, and J. Green, "Relation of automatically extracted formant trajectories with intelligibilty loss and speaking rate decline in amyotrophic lateral sclerosis," in *Proceedings of the 17th Annual Conference of the International Speech Communication Association (INTERSPEECH '16)*, San Francisco, USA, 2016.

[10] T. Bäckström, J. Lecomte, G. Fuchs, S. Disch, and C. Uhle, *Speech Coding with Code-Excited Linear Prediction*. Springer, 2017, to be published.

[11] J. Garofolo, L. Lamel, W. Fisher, J. Fiscus, D. Pallett, N. Dahlgren, and V. Zue, "TIMIT Acoustic-Phonetic Continuous Speech Corpus LDC93S1," Web download, Philadelphia: Linguistic Data Consortium, 1993.

[12] K. Mustafa and I. Bruce, "Robust formant tracking for continuous speech with speaker variability," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, 2006.

[13] T. Rossing, P. Wheeler, and F. Moore, *The Science of Sound*. Addison Wesley, 2002.