RECOVERING IMPLICIT PITCH CONTOURS FROM FORMANTS IN WHISPERED SPEECH

Pablo Pérez Zarazaga, Zofia Malisz
TMH, KTH Royal Institute of Technology, Stockholm
(pablopz,malisz}@kth.se

ABSTRACT

Whispered speech is characterised by a noise-like excitation that results in the lack of fundamental frequency. Considering that prosodic phenomena such as intonation are perceived through $f_0$ variation, the perception of whispered prosody is relatively difficult. At the same time, studies have shown that speakers do attempt to produce intonation when whispering and that prosodic variability is being transmitted, suggesting that intonation "survives" in whispered formant structure.

In this paper, we aim to estimate the way in which formant contours correlate with an "implicit" pitch contour in whisper, using a machine learning model. We propose a two-step method: using a parallel corpus, we first transform the whispered formants into their phonated equivalents using a denoising autoencoder. We then analyse the formant contours to predict phonated pitch contour variation. We observe that our method is effective in establishing a relationship between whispered and phonated formants and in uncovering implicit pitch contours in whisper.

Keywords: Whispered speech, Formant contours, Pitch contours, Intonation, Machine learning

1. INTRODUCTION

Pitch is an important feature in the encoding of prosody, and its variations over time are a defining characteristic of intonation. In whispered speech, fundamental frequency is not present in the signal as the glottis remains open and the vocal folds do not vibrate. In principle, this should prevent transmission of pitch and intonation via whisper.

However, several studies [1, 2, 3] have found that listeners perceive pitch and intonation effects in whispered speech. Similarly, it is evident that the whispered signal contains properties that offer cues to intonation.

Formant raising, particularly of F1, is a well-attested characteristic of whispered speech relative to phonated speech [4, 5, 6]. The higher formant frequency positions in whisper may be related to a more open configuration of the vocal tract similar to that found in e.g. Lombard speech [6, 4]. In our previous work, we showed that the jaw is more open in whispered, relative to phonated, vowels in Swedish, supporting the hypothesis that whisper is a form of hyperspeech. However, others have suggested that formant raising is related to whispered pitch perception: [3] showed that listeners were able to discriminate between low and high-pitched whispered vowels on the basis of higher F1 and F2. Additionally, raised formant frequencies seem to connect to prosody-related laryngeal activity [5]. [7] showed that in whispered speech production, laryngeal movements associated with pitch changes remain comparable to those in phonated speech. This indicates that prosodysrelated oral cavity shape modification takes place in whisper and hence is able to change the acoustics of the noise-like excitation travelling through the vocal tract - allowing for pitch inference in perception.

The harmonic relation between $f_0$ and formants has also been exploited in speech technology. In voice conversion systems, i.e. systems that reconstruct phonated speech from whispers, some solutions have tackled the problem of the missing $f_0$ by creating harmonic excitation for whispered speech using generative adversarial networks [8, 9]. [10] showed promising results by adding an artificial pitch model to formant structure.

In the present paper, our process runs somewhat in the opposite direction: we first aim to learn from a) the relationship between $f_0$ and formants in phonated speech and b) its connection to the spectral properties of whispered speech. To this end, we present a machine learning method that models the variation in $f_0$ present in phonated speech through the variation in whispered formant values. One goal is to use these relationships to uncover and understand the "implicit" pitch contour in whisper that allows for the perception of prosody in the absence of $f_0$. The other is to support technological

---

*This work was supported by the Swedish Research Council grant no. 2017-02861 “Multimodal encoding of prosodic prominence in voiced and whispered speech".
applications such as voice conversion.

2. METHOD

2.1. Data

In this work, we use the CHAINS dataset [11], which contains paired recordings in both phonated and whispered speech. CHAINS contains data from 20 male and 16 female speakers with Irish (28) and American (8) accents. The dataset consists of read sentences and text fragments in different speech modes such as natural phonated and whisper. We use a subset of CHAINS with paired phonated and whispered sentences and analyse only vocalic phonemes. We transcribed the phonated audio samples using a Wav2Vec2 model [12] and copied the transcriptions into the whispered equivalents. The data was then labelled using the Montreal Forced Aligner [13] with a pre-trained model for American English. We observed that function word reductions produced event sequence mismatches between phonated and whispered pairs. They were subsequently removed from the analysis yielding an exactly matched dataset of 7549 phonated and whispered vowel phonemes.

2.2. Feature extraction and preprocessing

We extracted $f_0$ and formant contours from vowel phones. $F_0$ values were extracted using Praat, with a time step of 0.01 s and using a pitch floor and ceiling of 75 and 300 Hz respectively. To extract the formant contours, we used the Burg method in Praat with a time step of 0.01 s, a 0.025 s long analysis window and the maximum number of formants set to 5. Pre-emphasis was applied over 50 Hz. The formant ceilings were initialised to 5000 Hz for male speakers and 5500 Hz for female ones and optimised using the Escudero’s method [14]. In order to normalise the length of each segment, we defined evenly distributed observation points every 10% of the total duration of the phone. The contour measurements corresponding to 0% and 100% were then discarded to avoid instabilities due to formant transitions between phones, resulting in 9 measurement points that define the pitch and formant contours. To reduce speaker variability in the formant contours, we normalised the data for each speaker following the Lobanov’s normalisation method [15], where the data presents zero mean and unit variance with respect to its corresponding speaker. Additionally, pitch contours are also normalised to zero mean and unit variance relative to each speaker.

2.3. Whispered pitch prediction

We propose a machine learning method to estimate the "implicit" $f_0$ contour in whisper from whispered formant structure. The two-step method is depicted in Figure 1. In the first step, we leverage the relationship between whispered and phonated formants residing in paired phones to provide a de-noised representation of the whispered formant frequencies. In the second step, we predict implicit, whispered $f_0$ contours by modelling the corresponding phonated $f_0$ with the denoised formant representations.
2.4. Formant denoising

We assume that whispered formant structure is strongly correlated to its phonated equivalent. With the difference that in whisper, the formant frequencies are raised and the spectral envelope peaks are flatter manifesting a noisier behaviour. The vowel spaces obtained from the current dataset are presented in Figure 2. A high variance in some $F_1$ values can be observed due to the scarce representation of some phones in the dataset. The effects of whispered speech are reflected as an increase in the values of $F_1$ and $F_2$ and a higher variance in whispered, relative to phonated, formant values.

Hence, we consider whispered formant contours as a noisy representation of phonated formants and propose a denoising strategy to "transform" whispered formant contours to phonated equivalents. For this purpose, we use an autoencoder [16] that learns higher-order statistical information from the formant contours. Using whispered-phonated formant pairs as input and output of the model, we aim to map the contours from whispered speech to the formant contours of the same phones in phonated speech.

The machine leaning method consists of a symmetric encoder-decoder structure. Three 1-D convolutional layers are used as filters over the time-dependent features (16 channels, kernel length of 3). These layers, plus a fully-connected layer, generate encoder embeddings (18-sample) from whispered formant contours. The decoder uses three transposed convolutional layers to invert these embeddings into phonated speech formants.

We have observed that most of the contours in the data do not present significant variations. This might introduce bias in our network towards a constant value in the formant and pitch contours. In order to maximise the similarity with the target contour, we choose a different loss function: the cosine distance between target and predicted contours, such that:

$$ L_{\text{cos}} = \frac{1}{N} \sum_{i=0}^{N} \frac{\hat{y}_i \cdot y_i}{||\hat{y}_i||_2 ||y_i||_2} $$

where $y_i$ and $\hat{y}_i$ stand for the target and denoised formant sequences of the $i$th sample and $\cdot$ represents the dot product between the two vectors.

The correlation coefficients for $F_1$, $F_2$ and $F_3$ are presented in Table 1. We can observe that, while the correlation between phonated and denoised formant contours is improved for $F_1$, $F_2$, the network degrades the correlation for $F_3$ values. Therefore, we chose to apply denoised $F_1$ and $F_2$, while the contour of $F_3$ is used unmodified in the next step.

2.5. Implicit $f_0$ prediction

The second step in the method is a sequence prediction model that estimates the "implicit" $f_0$ contour based on the denoised contours of $F_1$, $F_2$ and (unmodified) $F_3$. This works under the assumption that the $f_0$ contour in the phonated data is related to the one "implied" in the whispered data. Admittedly, modelling the exact values of a non-existing pitch signal from formant contours is a challenging task. Therefore, this model focuses on estimating the relative changes in the pitch contour that represent intonation variability.

We use a recurrent neural network (RNN), a deep learning method that is especially efficient in uncovering temporal dependencies within sequences. Considering the temporal dependencies in the formant and $f_0$ contours, a sequence-to-sequence model [17] should be particularly useful in modelling dependencies within and between dynamic contours. The input formant contours are processed by two bi-directional LSTM layers [18] with 4 hidden units, resulting in a sequence with 8-dimensional features. The output sequence is then generated with a fully-connected layer that maps the corresponding 8 features to the target pitch values.

Similarly to the formant denoising, we require a function that will maximise the similarity to the target sequence. The cosine distance can be seen as a normalised correlation, thus providing a better approximation to the sequence shape than mean squared error (MSE). Additionally, in order to also approximate the actual pitch values of the target, we will also include MSE in the loss function. The resulting loss function is then a combination of MSE and cosine distance:

$$ L_{\text{seq}} = \frac{1}{N} \sum_{i=0}^{N} |\hat{y}_i - y_i|^2 - \frac{1}{N} \sum_{i=0}^{N} \frac{\hat{y}_i \cdot y_i}{||\hat{y}_i||_2 ||y_i||_2} $$

Both the denoising and prediction models were trained over 300 epochs with an Adam optimiser and a learning rate of $10^{-4}$. Additionally, a recurrent dropout of 0.4 was added in the training loop to the LSTM layers of the $f_0$ prediction network.

3. RESULTS

First of all, our results show that we were able to successfully leverage the close relationship between whispered and phonated speech to denoise whispered formants. Table 1 shows that denoised whispered formant contours exhibit a higher correlation with their phonated counterparts. The greatest improvement can be seen in $F_1$, known to exhibit the greatest difference between phonated and
whispered speech [4, 5, 6]. In Figure 3, we present several examples of the denoised formant contours in which the denoising model acts as a smoothing function. It is evident that the cosine similarity loss function allows us to closely estimate formant frequency contours in stationary vowel intervals found in the CHAINS corpus. Where the network fails is in some extreme cases with high contour variability, which may be due to various factors such as formant tracking errors, residual influence of the flanking phones etc.

Second, the measured error and correlation values in the pitch prediction step are summarised in Table 2. The refined pitch contour (RPC) from [10] is used as baseline. Our model obtains a positive correlation (Mean $r = 0.56$, SD $= 0.3$) between predicted and target contours that allows to generally follow the target pitch. Some examples of the resulting mapping between predicted and phonauted contours can be observed in Figure 4. We see a correct direction of the trend in most cases. We also compare the absolute difference between the mean $f_0$ frequencies. Regarding comparisons to the RPC baseline, while the LSTM shows improvement in several individual cases, the average performance is similar to the baseline in correlation and error values. This leads us to believe that a more complex system such as LSTM would benefit from using additional input features supplying information beyond the minimalistic formant set.

The absolute error from the predicted $f_0$ presents an average value of 38 Hz, with a standard deviation of 24 Hz. The number of different speakers (36) in this data was high and this error can be explained by the impact of speaker variability. This estimation could be improved by considering speaker-dependent features or tuning the model specifically for each speaker.

### Table 1: Pearson’s correlation coefficient of whispered and denoised formant contours with respect to phonated ones.

<table>
<thead>
<tr>
<th>Formant</th>
<th>Whispered</th>
<th>Denoised</th>
</tr>
</thead>
<tbody>
<tr>
<td>$F_1$</td>
<td>0.19</td>
<td>0.34</td>
</tr>
<tr>
<td>$F_2$</td>
<td>0.44</td>
<td>0.47</td>
</tr>
<tr>
<td>$F_3$</td>
<td>0.26</td>
<td>0.16</td>
</tr>
</tbody>
</table>

### Table 2: Absolute error and correlation of implicit $f_0$ prediction from the refined pitch contour (RPC) estimation [10] and the proposed LSTM-based network with respect to phonated $f_0$.

<table>
<thead>
<tr>
<th>Method</th>
<th>RPC [10]</th>
<th>LSTM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Error (Hz)</td>
<td>Mean St. Dev.</td>
<td>Mean St. Dev.</td>
</tr>
<tr>
<td>Correlation</td>
<td>0.55</td>
<td>0.56</td>
</tr>
</tbody>
</table>

### Figure 3: Denoised formants from whispered contours with respect to original values.

### Figure 4: Predicted implicit $f_0$ contour values compared to the original contours.

**4. CONCLUSION**

In this work, we have presented a machine learning method that estimates implicit $f_0$ contours from formant contours in whispered vowels. Our models have shown that it is possible to uncover an implicit $f_0$ trajectory in whisper, via its phonauted equivalent, from $F_1$, $F_2$ and $F_3$ contours only.

In future lines of work, we will apply a speaker-dependent model. Our results show that the variability inherent in the 36 speakers modelled in this data has an influence on the $f_0$ prediction. Additionally, we would like to incorporate information on the phonauted phones and their context into the formant analysis and pitch estimation. The present models showed promising results even before this variability was taken into account and we expect that adding the additional features will decidedly improve performance.

This result will allow us to observe variations in perceived pitch and provide an improved analysis of intonation in whispered speech. The implementation of modern methods to uncover implicit pitch in whisper will also lead to developments in voice conversion systems.
5. REFERENCES


