On the Importance of Pure Prosody in the Perception of Speaker Identity

Elina E. Helander¹, Jani Nurminen²

¹Institute of Signal Processing, Tampere University of Technology, Finland
²Nokia Technology Platforms, Tampere, Finland

elina.helander@tut.fi, jani.k.nurminen@nokia.com

Abstract

Many of the current techniques and systems that deal with speaker identity do not regard detailed prosody as a crucial source of speaker-dependent information. The reasoning behind this relates to the common assumption that the F0 level and the spectral data carry all or almost all of the speaker-dependent information. But is this assumption really valid? We have investigated the importance of prosodic information in the perception of speaker identity by conducting a test where the listeners tried to identify people they know after hearing only delexicalized pure prosody signals. The findings presented in this paper show that even a very rough prosodic representation consisting only of a single sinusoid can contain information on speaker identity, giving motivation for the development and wider usage of techniques that better exploit the prosodic aspects.

Index Terms: prosody, speaker identity

1. Introduction

Voice conversion techniques aim at converting a speech signal to sound as if it was uttered by another speaker. This research topic has gained a lot of interest during recent years, e.g. [1], [2], [3], [4]. Most voice conversion systems published in the literature neglect detailed prosody modeling and concentrate mainly on spectral conversion. Fundamental frequency (F0) contours are usually modified in a very simple manner using e.g. Gaussian mixture modeling (GMM) based conversion or a simple mapping based on the mean F0 values and the variances of F0 [1], [3], [4]. The lack of wider interest towards more sophisticated prosody conversion techniques could originate from the fact that researchers have typically focused on the problem of changing some stranger’s voice into the voice of another stranger. In this kind of cases the detailed conversion of prosodic features may not be highly crucial. However, when thinking about the potential applications of voice conversion, it seems quite likely that the users would be interested in applications where they can convert some source voice into the voice of a person they know (friends, family, celebrities, etc.). In this scenario, the user is quite familiar with the person’s speaking style and our assumption is that the user might be quite likely to detect peculiarities in intonation patterns or tempo if prosody is not properly converted.

We have recently introduced a novel approach for converting prosody in voice conversion [5]. This new approach was evaluated in a listening test in connection with a complete voice conversion system [6] and it was found to clearly outperform the conventional GMM based F0 conversion scheme. In fact, the results presented in [5] indicated that the more accurate modeling of prosody increased the performance of voice conversion in two ways: the speaking style in the converted samples was found to be closer to the target, and the converted speech was found to be less robotic or monotonic. In the listening test, the target speaker was allowed to speak more freely than the source speaker from the prosodic point of view. While this is well in line with the potential use case where a text-to-speech (TTS) system acts as the source, it may also raise an important question: could the observed enhancement be caused by the increased expressiveness in the converted speech? Furthermore, can we be sure that the prosody contains person-dependent information that should be converted, other than the F0 level and the rough speaking style that could also be approximated using the variance of F0? The listeners judging the success of prosody conversion in [5] were not previously familiar with the target speaker and thus they evaluated more the speaking style than the real identity.

To further support our claims on the importance of prosody from the viewpoint of speaker identity, we show in this paper that people can identify themselves or persons they know on the basis of pure prosodic stimuli, although the average F0 levels of the different speakers are very close to each other. The pure prosody signals are obtained by estimating F0, duration and intensity information from the original signals and by creating a signal containing a single sinusoid whose frequency follows the F0 contour and whose amplitude follows the intensity contour of the original speech signal. This simplified pure prosody signal generation approach is very close to that proposed in [7]. The results, together with the earlier results presented in [5], give a clear indication that the prosodic aspects would deserve more attention in the voice conversion related research and development work.

This paper is organized as follows. In Section 2, some issues related to prosody and speaker identity are discussed. Section 3 describes the listening tests and the results obtained in the test. The discussions and the conclusions presented in Section 4 and Section 5, respectively, conclude this study.

2. On prosody and speaker identity

Prosody is a supra-segmental phenomenon that is not conveyed through a single phonetic segment but through larger units like words, sentences, utterances or even paragraphs. A vast amount of research has been devoted to prosody in relation to text-to-speech systems during the previous decades, especially related to intonation that is considered perhaps the most important aspect of prosody. Intonation can have linguistic, paralinguistic or extralinguistic functions [8]. Linguistic functions comprise of the morphological and lexical levels of phrase, as well as of discourse and dialogue levels. Emotions and mental states are categorized as paralinguistic. Extralinguistic factors like age and sex are personal and physical characteristics. These different linguistic sources of information are resolved in the ear via perceptual processing. Based on the formant frequencies and...
the pitch, the listener usually forms an estimate if the speaker is a male, female or a child, even if he/she does not know the speaker.

2.1. Prosody and speaker identity in the literature

It is widely accepted that prosody plays an important role in naturalness of speech. In text-to-speech systems, the aim is to generate a “good” and acceptable prosody. However, in natural speech there exists a great deal of inter-speaker and intra-speaker variations in prosody [9]. An interesting experiment was made in [10] where five speakers of the same sex and age were presented to familiar listeners in different forms of natural and synthetic stimuli. It was shown that the speakers were highly identifiable on the basis of their fundamental frequency characteristic even when the spectrum was generated artificially. A slightly similar experiment was carried out in [11] where the spectral information related to speaker individuality was hidden using the average envelopes of all speakers. The pitch contours were shown to make a difference from the viewpoint of speaker individuality.

Many voice conversion systems transform the F0 contours through simple scaling between target’s and source’s mean F0 value and the standard deviation of F0. Some experiments with more detailed pitch conversion have been reported [2],[12],[13] but compared to the total number of voice conversion papers published so far, the topic of more detailed F0 modeling in voice conversion has not received much attention.

In addition to F0 contours, another important aspect of prosody are the durational issues: the durations of segments and pauses that are usually referred to as speech tempo. In voice conversion, the durations are usually modeled by simple scaling of the target’s and source’s utterance lengths. However, it was shown in [14] that there are inter-speaker differences between sound classes. This will not favor uniform speaking rate modifications. Rate modifications, global or local, are not entirely linear due to the properties of human speech production, but affect vocalic segments more than consonantal segments [9]. In [4], the durational modifications are separated into three distinct classes, silent, unvoiced, and voiced. It was claimed that more detailed modeling, i.e. phoneme level, would be too inaccurate but in [1] context-dependent duration difference modeling between speakers in phone or triphone level was suggested.

When looking outside the topic of voice conversion, most speaker recognition systems also ignore the prosodic features and concentrate only on the short-term spectral features. However, the addition of short-term [15] or long-term [16] prosodic information has been shown to improve the performance. Related to speaker identification, an interesting finding was made in [17] where an imitator was asked to mimic two subjects. Human listeners did not recognize the difference between the imitator and the real speaker nearly as well as a speaker verification system that did not use prosodic information. From the viewpoint of voice conversion, this supports the assumption that it could be possible to convince a human listener with even a slightly imperfect spectral conversion as long as the prosody is converted more accurately.

2.2. Some experimental findings

We have studied the prosodic features in the speech of different speakers and found that there are clear speaker-dependencies in speech signals that should be modeled for example in voice conversion. An example case including F0 contours from two speakers uttering the same sentence is shown in Figure 1. The straight lines represent the measured F0 contours and the dashed lines indicate the transformed F0 with the second speaker’s mean and variance applied to the first signal and vice versa. In the transformed signals, the different durations were taken into account by applying dynamic time warping to the speech signals. It can be observed that there are similarities due to the same context but some clear speaker-specific differences can also be found. Furthermore, it is also possible to see that the mean and variance based approach does not achieve very good results.

Further experimental findings were presented in [5]. In that paper, we proposed a new prosody conversion scheme that uses a syllable-based prosodic codebook. The selection from the codebook employs not only the source contour but also linguistic information and segmental durations with the aid of a trained classification and regression tree. The new method was included as a part of the voice conversion system originally presented in [6] and it was evaluated in a listening test. The results clearly indicated that more detailed prosody conversion enhanced the quality of the converted samples when compared to the conventional GMM based approach: it increased the naturalness (or made the output sound less robotic or monotonous) and the identity mapping was improved through the better modeling of the person-dependent speaking style.

Even after these findings and good results, one important question remains. Does our prosody conversion approach improve the results through better prosody modeling in general or does the prosody really contain some person-dependent features that can be modeled in voice conversion? The rest of this paper demonstrates through a listening test that the prosody indeed contains person-dependent information.

3. Listening tests

To get more evidence on the speaker-dependencies in prosody, a listening test involving pure prosodic signals was carried out. This approach was chosen to be able to really focus on the
prosodic aspects and to avoid any influence from other features in speech signals or from the voice conversion system.

3.1. Test arrangement

Several neutrally spoken sentences from 16 different native Finnish speakers were recorded in a quiet room. The recorded signals were analyzed automatically to obtain F0 and energy contours and these contours were used in the generation of pure prosody signals. In these signals, the voiced regions contain a single sinusoid whose frequency follows the F0 contour and whose amplitude is in line with the energy contour, while the unvoiced and silent regions are represented as silence. An example including a speech signal and the corresponding pure prosody signal is shown in Figure 2.

In total 14 native Finnish listeners participated in the test. Each listener heard sinewave signals generated from speech spoken by people she/he knows (possibly including her/himself) in random order and was asked to identify the speaker. For some of the sinewave signals, the listener was shown the textual form of the corresponding sentence while for other sinewaves no supplementary information was given. Before starting the actual test, each listener was given a chance to get familiar with the sinewave representation. In this training phase, each listener was able to listen to both the original recorded form and the corresponding sinewave form of some example sentences as many times as she/he wanted. One sentence from each speaker was included in this training material.

Many of the listeners also served as a speaker and vice versa since we also wanted to examine how well people recognize themselves from the sinewave based prosodic representation. The speakers were grouped into three categories based on the gender and age: in total there were 6 female, 8 male and 2 child speakers. The speakers were also further divided into groups of people who knew each other. Each listener heard only samples from her/himself or from speakers that she/he knows from some context: either as a family member or a close relative, as a friend, or as a colleague. Each listener was explicitly given the list of speakers that could appear in the samples. If we assume that the listener can always recognize the correct category (female, male or child voice), there were always either two or three alternatives to choose from. This assumption was found valid with only a couple of exceptions that were most likely caused by unfamiliarity with the sinewave representation. The two or three alternatives were chosen using two criteria: the listener has to know the speakers and their F0 levels were very close to each other. This applied in all the cases except for the children.

3.2. Results

Table 1 illustrates the overall result obtained from the whole test. The notations “From 2” and “From 3” denote the number of alternatives the listeners can choose from provided that the answer falls in the correct category (male, female, or child). As can be seen from the average identification rates obtained from a total of 524 answers, it can be concluded that even the very rough representation of pure prosody including only a single sinusoid contains speaker-dependent information. Even though the sinewave representation was a rather odd and unfamiliar representation for the listeners, it still helped significantly in the speaker identification. For comparison, the table also shows the average rate that would be achieved by guessing (with the assumption that the listener can always recognize the speaker category correctly). The difference between the test result and the average guessing rate is statistically significant based on the fact that the upper bound of the confidence interval in the case of guessing is lower than the lower confidence interval of the test results. We also observed that there was no difference on how listeners recognized themselves compared to others.

The potential influence of having the textual version of the sinewave signal available is analyzed in Table 2. Indeed, the identification rate seems to slightly improve when the listener knows the sentence that is spoken but this result is not statistically significant considering the confidence intervals.

4. Discussion

Although the results are favorable as such, some further issues should be pointed out. First, the language used in the experiments was Finnish that is regarded as a rather monotonous language in terms of F0 contours. This was especially true in the case of some male speakers. Additional results using some other language could reveal even more evident findings. Moreover, we also observed that the presence of the microphone and
Table 2: The effect of having the text available

<table>
<thead>
<tr>
<th>From 2</th>
<th>From 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>No. of sentences without text</td>
<td>209</td>
</tr>
<tr>
<td>Recog. without text</td>
<td>66 %</td>
</tr>
<tr>
<td>Upper conf. interval</td>
<td>72 %</td>
</tr>
<tr>
<td>No. of sentences with text</td>
<td>75</td>
</tr>
<tr>
<td>Recog. with text</td>
<td>81 %</td>
</tr>
<tr>
<td>Lower conf. interval</td>
<td>71 %</td>
</tr>
</tbody>
</table>

the usage of pre-defined sentences did not help in capturing natural prosodics that would make the identification easier.

It was also found that special attention should be paid to the planning of the text materials when studying the prosodic effects. This also applies more widely e.g. to voice conversion since the planning of a limited number of sentences that could capture the spectral, the prosodic and the speaking style related individualities poses a real challenge. Voice conversion prompt sheet planning is not covered well in the literature.

We believe that the results were affected by the fact that pure prosodic signals were strange and unfamiliar to the listeners. Some people adapted to listening to the sinewave representations faster than others but further training might have helped all the listeners. In fact, many listeners commented that more practice with the sinewave signals could have improved their performance substantially. The listeners found dealing with pure prosody signals more comfortable and easier when they had the opportunity to see the texts. However, the general recognition rate was not significantly improved, as shown in Table 2.

Finally, it should also be noted that the F0 levels of the speakers in the same category (male, female or child voice) were often very close to each other. The listeners had to identify speakers whose F0 levels were as little as 3 Hz apart.

5. Conclusions

In this study, we have discussed the importance of prosody in the perception of speaker identity. We conducted a listening test in which 14 listeners were asked to identify speakers based on simplified prosodic signals. Each utterance consisted only of a single sinewave following the F0 and energy contours of the corresponding recorded sentence. It was shown that it is possible to identify familiar people on the basis of pure prosody. From the viewpoint of voice conversion, the result, together with the results we have presented earlier in [5], highlights the potential advantage that may be gained in voice conversion through proper treatment of prosody. This will presumably be especially true in relation to commercial voice conversion applications since the consumers are likely to be interested in having a voice converted to the voice of a person whose prosodic peculiarities they know very well.

6. Acknowledgements

This work was partially supported by the Academy of Finland project No. 5213462 (Finnish Centre of Excellence program 2006-2011). Moreover, this work has partially been funded by the European Union under the integrated project TC-STAR - Technology and Corpora for Speech-to-Speech Translation - (IST-2002-FP6-506738, http://www.tc-star.org).

7. References