Robust intonation pattern classification in human robot interaction

Martin Heckmann1, Kazuhiro Nakadai2, Hirofumi Nakajima2,3

1Honda Research Institute Europe GmbH, D-63073 Offenbach/Main, Germany
2Honda Research Institute Japan Co. Ltd., Wako-shi, Saitama 351-0188, Japan
3Department of Computer Science, Faculty of Informatics, Kogakuin University, Japan
martin.heckmann@honda-ri.de, nakadai@jp.honda-ri.com, nakajima@jp.honda-ri.com

Abstract

We present a system for the classification of intonation patterns in human robot interaction. The system distinguishes questions from other types of utterances and can deal with additional reverberations, background noise, as well as music interfering with the speech signal. The main building blocks of our system are a multi channel source separation, robust fundamental frequency extraction and tracking, segmentation of the speech signal, and classification of the fundamental frequency pattern of the last speech segment. We evaluate the system with Japanese sentences which are ambiguous without intonation information in a realistic human robot interaction scenario. Despite the challenging task our system is able to classify the intonation pattern with good accuracy. With several experiments we evaluate the contribution of the different aspects of our system.

Index Terms: Intonation pattern classification, human robot interaction, source separation, pitch tracking

1. Introduction

It is well known that prosodic information plays an important role for human-human communication. Nevertheless, it is still rarely used in human-machine interaction [1, 2, 3, 4]. Reasons for this are a limited understanding of the prosodic structure of speech, ambiguities in the prosodic cues, and difficulties of robustly extracting the relevant cues.

In human robot communication one important additional challenge is to cope with acoustically adverse environments. To render the communication natural it is required that the robot perceives its acoustical environment not via a headset worn by the user but via microphones mounted on the robot. As a consequence the speech signals acquired by the robot are impaired by room reflections and additional sound sources present in the room, thereby further complicating the extraction of prosodic cues. Nevertheless, some steps towards integrating them into human-robot interaction have been made [5, 6].

In this paper we present a system which is able to distinguish different intonation patterns form utterances directed to a robot. In contrast to other approaches we do not use any lexical information. Our focus is to investigate how reliable the relevant acoustic cue, i.e. the fundamental frequency variation in the speech signal, can be extracted in a realistic interaction scenario with additional noise sources present. As the fundamental frequency alone does not yield reliable cues to distinguish statements, affirmations, and denials we restrict our system to distinguish questions from these classes. Questions typically show a rising fundamental frequency on the final speech segment [7]. Consequently we determine the final speech segment and classify the found fundamental frequency pattern.

Our system combines different building blocks to obtain a robust extraction and classification of the relevant fundamental frequency contours (compare Fig. 1). The first is a multi channel source separation which enhances the signal. The second step is an algorithm for fundamental frequency extraction, whose percept is called pitch, which takes inspirations from models of human pitch perception. The next step is the deployment of a Bayesian tracking algorithm on the resulting histograms. A voicing detection serves to determine for which segments the pitch has to be evaluated. For reliable Voice Activity Detection (VAD) we apply a post filter on the speech signal after source separation and add a further component for the elimination of crosstalk. On this signal we perform VAD. Using an energy based syllabification we determine the final and for our task relevant last segment of the utterance. We then classify the pitch movement in this final segment by comparing it to several reference patterns via Dynamic Time Warping (DTW).

In the following we will detail the building blocks of the proposed system for intonation classification. After this we will give an overview on the human-robot interaction scenario in which we tested our algorithm. The presentation of the results and their discussion will conclude the paper.

2. Geometric-constrained High-order Decorrelation-based Source Separation

We use Geometric-constrained High-order Decorrelation-based Source Separation (GHDSS) for sound source separation, mainly to suppress directional noise sources. It builds upon Decorrelation-based Source Separation (DSS) using Independent Component Analysis (ICA) in the frequency domain and includes Geometric-constraints to overcome permutation and scaling problems [8]. Furthermore, it features an adaptive step-size control to cope with changes in the environment.

After the separation for further processing the source from the frontal direction is chosen and transformed back into the time domain via application of the Inverse Fast Fourier Transform (IFFT).

3. Voicing Calculation

The information on the voicing of a segment is needed to determine if pitch has to be evaluated for this segment. We consider a segment voiced if the normalized cross correlation $q_{NCCF}(k)$, given by

$$q_{NCCF}(k, \kappa) = \frac{1}{N} \sum_{j=k}^{k+N} \frac{r(j)r(j + \kappa)}{\sqrt{e(k)e(k + \kappa)}},$$

(1)
where \( r(k) \) is the signal at time index \( k \) and \( e(k) \) its corresponding energy, is larger than a given threshold \( t_e \) [9].

### 4. Pitch Estimation

The algorithm we apply for pitch extraction is inspired by human pitch perception models [10] and relies on the calculation of a histogram of zero crossing distances and a subsequent inhibition of side peaks resulting from harmonics and sub-harmonics of the true fundamental frequency [11, 12].

First we split the signal into different frequency channels via the application of a Gammaton filter bank. Next we scan through possible fundamental frequency hypotheses \( f_0^h \) and set up a comb filter with teeth at the location of the harmonics \( l \cdot f_0^h \). By comparing found patterns with expected patterns from harmonics and subharmonics of the current hypothesis \( f_0^h \) we are able to suppress spurious side peaks at these harmonics and subharmonics. Summing up all hypotheses we obtain a histogram \( h \) of likelihoods for the different hypotheses [11, 12].

On the histogram \( h \) we apply a tracking algorithm based on Bayesian filtering [13, 14]. It sequentially integrates in the estimation of the state \( x_k \) at time \( k \) information from a model on the pitch dynamics \( p(x_k|x_{k-1}) \) and observations from the pitch histogram \( p(z_k|x_k) \). A subsequent backward pass, termed Bayesian smoothing, integrates information on future observations to improve performance [13].

### 5. Voice Activity Detection

For the proposed intonation classification we use the pitch movement of the final speech segment. Therefore, the precise determination of the end of the speech segment is crucial. To obtain this we apply a three stage Voice Activity Detection (VAD) processes. The first two stages further enhance the signal resulting from the GHDSS and the third one performs the actual VAD.

#### 5.1. Histogram-based Recursive Level Estimation (HRLE)

To further enhance the speech signal after the GHDSS we use Histogram-based Recursive Level Estimation (HRLE) [15]. Since HRLE uses recursive averages it calculates a time-varying histogram in real-time. Therefore, the noise level estimation smoothly and quickly adapts to the environmental changes.

#### 5.2. Crosstalk Suppression (CTS)

After application of the GHDSS we have access to two signal streams: the separated speech signal, \( y_s(k) \) and the separated music signal \( y_m(k) \). To minimize crosstalk between these two signals during Voice Activity Detection (VAD) we determine the regions in the spectrograms \( S_j \) of these two signals where the energy of either of the two streams is higher than the other. From this we calculate an enhanced speech spectrogram

\[
S_{SEnh}(k, \omega) = \begin{cases} S_0(k, \omega) & \text{if } S_0(k, \omega) > S_M(k, \omega) \\ 0 & \text{otherwise} \end{cases}
\]

which contains only those regions in the speech stream where speech dominates. The signal energy \( SEnh(k) \) is then obtained by summing \( S_{SEnh}(k, \omega) \) over all frequencies.

#### 5.3. Final VAD

Prior to the GHDSS we already performed a coarse Voice Activity Detection (VAD) with the MUSIC algorithm as implemented in [16]. Based on this segmentation we calculate the mean energy of the enhanced speech signal \( SEnh(k) \). Values larger than 60% of this value are taken as speech activity. Applying a median filter of length 100 ms on this signal fills gaps shorter than 100 ms. A second median filter of length 200 ms on one hand fills the gaps further but more importantly removes segments shorter than 200 ms.

### 6. Intonation Classification

We use the pitch track resulting form the previous step to identify the intonation pattern.

For doing so we first have to identify the final segment of the speech signal on which the classification should be based. More precisely, based on the VAD we determine the final segment, and classify it as belonging to one of four different patterns.

#### 6.1. Segmentation

To find the last segment in the speech segment detected by the VAD we use a syllabification algorithm [17]. It is based on the algorithm described in [18] and only uses the signal energy to find the syllable boundaries. This results in general in an over segmentation which is counterbalanced by following post-processing which yields reasonable estimates of the final speech segment. If the found segment, i.e. syllable, is shorter than 150 ms we add further syllables until they span at least 150 ms. A final segment longer than 300 ms is cut to 300 ms. As pitch is undefined in unvoiced regions we linearly interpolate between the surrounding voiced segments for all unvoiced regions.

#### 6.2. Classification

For classification we compare the pitch movement in the final speech segment \( s_F \) to four different prototypes \( s_F^{(i)} \) depicted in Fig. 2. These prototypical pitch movements aim to cover rising pitch movements in questions (r: a rising final segment, p: falling from a higher level with a final rise) and pitch movements found in the other classes (f: a flat final segment, d: a falling final segment). The prototypes have equal length and a mean of zero.

We apply Dynamic Time Warping (DTW) [19] to compare the final segment \( s_F \) to these prototypes. For doing so we also subtracted the mean pitch value from \( s_F \). The prototype \( s_F^{(i)} \) yielding the smallest distance is selected as the matching one.
In addition to this we also added music to the signals. The music signal was not present during the recording but added artificially by convolving the music signal with the transfer function measured for a direction of 60° from the front to the room. In addition we also made recordings from a headset the users were wearing. We used this signal only to benchmark our intonation classification system. The signals recorded on the robot are already impaired by the reverberations from the room and the background noise present in the room. In addition to this we also added music to the signals. The music signal was not present during the recordings but added artificially by convolving the music signal with a transfer function measured for a direction of 60° from the left of the robot. The music signal had approximately the same power as the speech signal, i.e. the Signal-to-Noise Ratio was around 0 dB.

During the interaction 4 male users were uttering 40 different Japanese sentences which are ambiguous without intonation information. They were uttering them with different intonation patterns, e.g. TA NO SHI KA Q TA (I enjoyed it./Did you enjoy it?). As intonation patterns we used 152 questions, 148 affirmations, and 104 denials, yielding 404 utterances. For the classification we combined affirmation and denial into one class such that the distinction is only between 152 questions and 252 items in the remaining class.

The signals are recorded with 16 kHz sampling rate. We use a Gammatone filter bank with 100 channels with center frequencies from 50…5000 Hz for the pitch extraction. The maximal pitch value was set to 500 Hz and the Bayesian smoothing, a part of the Bayesian tracking, was performed on 200 ms.

To assess the contribution of the different parts of our system we performed different tests. First we evaluated the intonation classification from the microphone closest to the speakers (referred to as BestMic). In a second test we use the same signal but perform the VAD calculation on the signal at the output of the GH DDS (referred to as BestMicGH DDS VAD). This highlights the importance of the segmentation. The comparison of the BestMic results with those obtained after the source separation via GH DDS show the contribution of the GH DDS. In a further test we added the HRLE post filter mentioned in Sec. 5.1 and the cross talk suppression mentioned in Sec. 5.2 to the VAD calculation (referred to as GH DDS+PostF/i o r a n t ) . Thereby we can determine the impact of this further enhancement step. Finally we also use the headset signal (referred to as Headset). These results allow us to delineate the impact of the adverse acoustical environment we face in realistic human-robot interaction.

For all cases mentioned above we performed the pitch tracking not only with the algorithm described in Sec. 4 but also with two publicly available and commonly used pitch tracking frameworks. These are get f0 from ESPS in the implementation of the WaveSurfer toolkit [20, 9] and praat [21]. Both frameworks are based on an autocorrelation calculated from the full-band signal.

In Fig. 3 the spectrograms for the BestMic, GH DDS and Headset case are shown. These results illustrate some of the difficulties encountered during the intonation classification. Despite the distortions in the signal in this example the voice activity detection and the pitch tracking is accurate in all three cases.
Table 1: Classification error rates in %.

<table>
<thead>
<tr>
<th></th>
<th>BestMic</th>
<th>BestMic</th>
<th>GHDS</th>
<th>GHDS</th>
<th>Headset</th>
</tr>
</thead>
<tbody>
<tr>
<td>GHDS VAD</td>
<td>38</td>
<td>29</td>
<td>28</td>
<td>26</td>
<td>20</td>
</tr>
<tr>
<td>praat</td>
<td>38</td>
<td>28</td>
<td>27</td>
<td>25</td>
<td>13</td>
</tr>
<tr>
<td>proposed</td>
<td>42</td>
<td>25</td>
<td>23</td>
<td>16</td>
<td>10</td>
</tr>
</tbody>
</table>

We obtain good intonation classification results on the clean signal. Nevertheless, some errors are present. The detailed analysis above for the headset case illustrated that they reflect the general difficulty of the precise extraction of the fundamental frequency as well as voicing information and the ambiguous nature of the pitch movements in respect to speech acts.

We could show that the classification results don’t deteriorate too much in an acoustically challenging environment mainly due to the source separation via GHDS, enhanced VAD, and the robust pitch extraction. Hereby the correct segmentation of the signal plays a crucial role. It is worth noting that when using the proposed source separation and the proposed pitch extraction the results on the noisy signals are almost as good as those obtained by applying the standard pitch extraction algorithms on the clean signal.

The approach we followed for only classifying the final pitch movement of the utterance is certainly too limited. To obtain better results additional features are required. On one hand it will be necessary to evaluate the pitch contour of the whole utterance, especially relating the last segment to the mean pitch value of a speaker. Furthermore, other cues to intonation than pitch have to be taken into account. This comprises e.g. the energy profile and the lengthening of the syllables. Nevertheless, we think that the results we obtain are suited such that the system we propose can be used to improve dialog act classification in human robot interaction and thereby serve as an additional cue to improve human robot communication.

9. Conclusion

We want to thank Lars Schillingmann for providing us the syllabification algorithm.

10. References