Robust intonation pattern classification in human robot interaction

Martin Heckmann, Kazuhiro Nakadai

2011

Preprint:

This is an accepted article published in Proc. INTERSPEECH. The final authenticated version is available online at: https://doi.org/[DOI not available]

Robust intonation pattern classification in human robot interaction

Martin Heckmann[1]*, Kazuhiro Nakadai*[2]*, Hirofumi Nakajima*[2]

[1]Honda Research Institute Europe GmbH, D-63073 Offenbach/Main, Germany
[2]Honda Research Institute Japan Co. Ltd., Wako-shi, Saitama 351-0188, 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. Distortions present in the speech signal are room reverberations, background noise, and a music source at 60°. 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 $q_{\mathrm{NCCF}}(k)$, given by

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

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

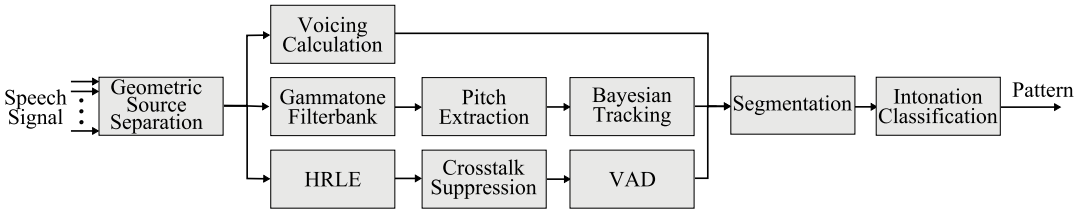


Figure 1: System overview

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 form 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' and set up a comb filter with teeth at the location of the harmonics $l \cdot f_0'$. By comparing found patterns with expected patterns from harmonics and subharmonics of the current hypothesis f_0' we are able to suppress spurious side peaks at these harmonics and subharmonics. Summing up all hypotheses we obtain a histogram $\mathbf{h}$ of likelihoods for the different hypotheses [11, 12].

On the histogram $\mathbf{h}$ 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, $y_{\mathrm{S}}(k)$ and the separated music signal $y_{\mathrm{M}}(k)$. To minimize crosstalk between these two signals during *Voice Activity Detection (VAD)* we determine the regions in the spectrograms $S_{[]}$ 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

$$S_{\mathrm{SEnh}}(k,\omega) = \begin{cases} S_{\mathrm{S}}(k,\omega) & \text{if} \quad S_{\mathrm{S}}(k,\omega) > S_{\mathrm{M}}(k,\omega) \\ 0 & \text{otherwise} \end{cases} \quad (2)$$

which contains only those regions in the speech stream where speech dominates. The signal energy $e_{\mathrm{SEnh}}(k)$ is then obtained by summing $S_{\mathrm{SEnh}}(k,\omega)$ 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 $\bar{e}_{\mathrm{SEnh}}(k)$. Values larger than 60% of this value are taken as speech activity. Applying a median filter of length $100\,\mathrm{ms}$ on this signal fills gaps shorter than $100\,\mathrm{ms}$. A second median filter of length $200\,\mathrm{ms}$ on one hand fills the gaps further but more importantly removes segments shorter than $200\,\mathrm{ms}$.

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.

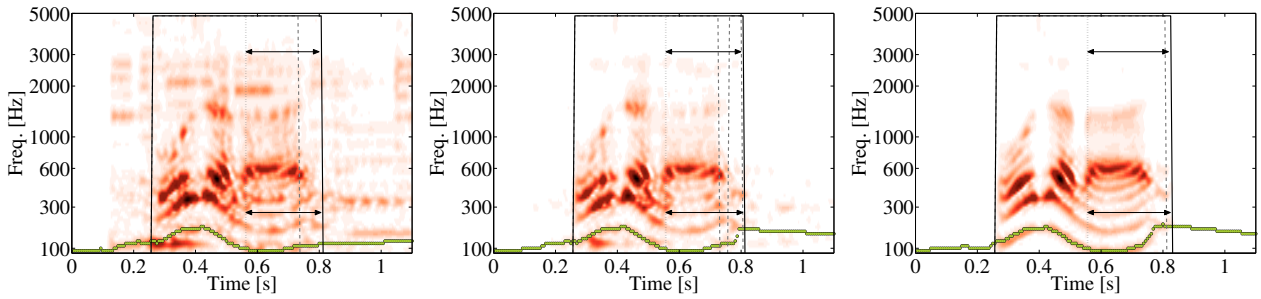
6.1. Segmentation

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 $150\,\mathrm{ms}$ we add further syllables until they span at least $150\,\mathrm{ms}$. A final segment longer than $300\,\mathrm{ms}$ is cut to $300\,\mathrm{ms}$. As pitch is undefined in unvoiced regions we linearly interpolate between the surrounding voiced segments for all unvoiced regions.

6.2. Classification

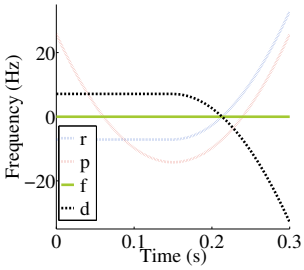
For classification we compare the pitch movement in the final speech segment s_F to four different prototypes $s_P^{(i)}$ depicted in Fig. 2. These prototypical pitch movements aim to cover rising pitch movements in questions (r: a rising final segment, p: falling from a higher level with a final rise) 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 $s_P^{(i)}$ yielding the smallest distance is selected as the matching one.



(a) (b) (c)

Figure 3: Spectrograms of one utterance for (a) the BestMic, (b) the GHDSS, and (c) the Headset case are shown. The extracted pitch contour is visualized in green and the detected speech region with the black dashed curve. Grey dashed lines indicate a transition from a voiced to an unvoiced region. The arrows and the grey dotted line indicate the final segment used for the classification of the pitch movement. Note that in the unvoiced regions pitch is interpolated.



(a)

(b)

Figure 2: (a) prototypical pitch movements used for classification, (b) our robotics platform Haerbo.

7. Evaluation

We asses the performance of our intonation classification system in a human-robot interaction scenario. Different people spoke to our robotics platform Hearbo, a wheeled platform with a humanoid upper body (compare Fig. 2). Hearbo has a height of 120 cm. In total it has 34 degrees-of-freedom and features different sensors. It perceives its acoustical environment via an 8 channel MEMS-based microphone array mounted around the top of its head. The sound acquisition and the GHDSS run on the robot in the open-sourced real time robot audition software HARK (HRI-JP Audition for Robots with Kyoto university) [16]. The remainder of the processing presented here is performed off-board and off-line even though the pitch extraction also runs in an online system [14]. We use the Robot Operating System (ROS) developed by Willow Garage, Inc. to control Hearbo.

The interactions took place in a 4 m $\times$ 7 m room with RT_{20} = 300 ms[1]. The users were standing at a natural interaction distance of $\approx$ 1.5 m and talking at an angle of 0° from the front to Hearbo. In addition we also made recordings from a headset the users were wearing. We used this signal only to benchmark our 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. Thereby 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 asses 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*)[2]. In a second test we use the same signal but perform the VAD calculation on the signal at the output of the GHDSS (referred to as BestMic$_{\mathrm{GHDSS\,VAD}}$). This highlights the importance of the segmentation. The comparison of the BestMic results with those obtained after the source separation via GHDSS show the contribution of the GHDSS. 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 GHDSS$_{+\mathrm{PostProc.}}$). 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, GHDSS 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.

[1] RT_{20} is better suited for measurements in noisy environments. It gives the decay measured at 20 dB extrapolated to 60 dB decay

[2] This is in fact a virtual microphone as it also includes the contribution of the simulated interfering music signal.

Table 1: Classification error rates in %.

	BestMic	BestMic GHDSS VAD	GHDSS	GHDSS +Post Proc.	Headset
praat	38	29	28	26	20
get_f0	38	28	27	25	13
proposed	42	25	23	16	10

However, the voicing detection is notably impaired. As a consequence also the pitch extraction is impaired as pitch tracks are only evaluated in voiced regions. This can be seen by comparing the final part of the BestMic case with the other two cases.

In Table 1 the classification results for the different cases are given. In the Headset case we obtain 40 (i. e. 10%) errors when using the proposed pitch tracking. Of these are 16 due to erroneous pitch tracking, 10 due to wrong voicing decision, 2 resulting from wrong segmentation, and in 12 cases our prototypical pitch patterns do not match the data (e. g. some negations have a rising pitch at the end). One can see that the degradation from the Headset condition to the condition without further processing (BestMic) is very significant and performance is close to chance level (50%). Supplying a better speech segmentation by using the GHDSS for VAD improves performance a lot. Separating the signals via GHDSS before classification further improves the performance. Finally, incorporating the post processing steps ($\text{GHDSS}_{+\text{PostProc.}}$) additionally reduces the error rates. When looking on the different pitch extraction algorithms one can see that the pitch extraction we proposed performs better than either praat or get_f0 in all cases tested ,except for the BestMic case without the VAD from GHDSS. In fact our pitch extraction performs on the unprocessed but correctly segmented signal ($\text{BestMic}_{\text{GHDSS VAD}}$) already better than the other two on the signal after source separation.

8. Conclusion

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 GHDSS, 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 of 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. Acknowledgments

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

10. References

[1] J. Ang, Y. Liu, and E. Shriberg, "Automatic dialog act segmentation and classification in multiparty meetings," in *Proc. ICASSP*. IEEE, 2005, vol. 1, pp. 1061–1064.

[2] V.K. Rangarajan Sridhar, S. Bangalore, and S. Narayanan, "Combining lexical, syntactic and prosodic cues for improved online dialog act tagging," *Computer Speech & Language*, vol. 23, no. 4, pp. 407–422, 2009.

[3] E. Nöth, A. Batliner, A. Kießling, R. Kompe, and H. Niemann, "Verbmobil: The use of prosody in the linguistic components of a speech understanding system," *IEEE Trans. Speech and Audio Proc.*, vol. 8, no. 5, pp. 519–532, 2000.

[4] M.Q. Wang and J. Hirschberg, "Automatic classification of intonational phrase boundaries," *Computer Speech & Language*, vol. 6, no. 2, pp. 175–196, 1992.

[5] S. Fujie, K. Fukushima, and T. Kobayashi, "A conversation robot with back-channel feedback function based on linguistic and nonlinguistic information," in *Proc. Int. Conf. on Autonomous Robots and Agents (ICARA)*, 2004, pp. 379–384.

[6] C. Breazeal and L. Aryananda, "Recognition of affective communicative intent in robot-directed speech," *Autonomous Robots*, vol. 12, no. 1, pp. 83–104, 2002.

[7] D. Hirst and A. Di Cristo, *Intonation systems: A survey of twenty languages*, chapter A survey of intonation systems, pp. 1–44, Cambridge University Press, 1998.

[8] H. Nakajima, K. Nakadai, Y. Hasegawa, and H. Tsujino, "Blind source separation with parameter-free adaptive step-size method for robot audition," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 18, no. 6, pp. 1476–1485, 2010.

[9] K. Sjölander and J. Beskow, "Wavesurfer-an open source speech tool," in *Sixth Int. Conf. on Spoken Lang. Proc. (ICSLP)*, 2000.

[10] A. de Cheveigne, "Pitch perception models," in *Pitch*, C. Plack and A. Oxenham, Eds. Springer, Cambridge, U.K., 2004.

[11] M. Heckmann, F. Joublin, and K. Nakadai, "Pitch extraction in human-robot interaction," in *Proc. IEEE/RSJ Int. Conf. on Intell. Robots and Systems (IROS)*, Taipei, Taiwan, 2010.

[12] M. Heckmann, F. Joublin, and C. Goerick, "Combining rate and place information for robust pitch extraction," in *Proc. INTERSPEECH*, Antwerp, 2007, pp. 2765–2768.

[13] C. Gläser, M. Heckmann, F. Joublin, and C. Goerick, "Combining auditory preprocessing and bayesian estimation for robust formant tracking," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 2, pp. 224–236, 2010.

[14] M. Heckmann, C. Gläser, M. Vaz, T. Rodemann, F. Joublin, and C. Goerick, "Listen to the parrot: Demonstrating the quality of online pitch and formant extraction via feature-based resynthesis," in *Proc. IEEE/RSJ Int. Conf. on Intell. Robots and Systems (IROS)*, Nice, 2008.

[15] H. Nakajima, G. Ince, K. Nakadai, and Y. Hasegawa, "An easily-configurable robot audition system using Histogram-based Recursive Level Estimation," in *Proc IEEE/RSJ Int. Conf. on Intell. Robots and Intell. Syst. (IROS)*, 2010, pp. 958–963.

[16] K. Nakadai, H.G. Okuno, H. Nakajima, Y. Hasegawa, and H. Tsujino, "An Open Source Software System For Robot Audition HARK and Its Evaluation," in *Proc. IEEE-RAS Int. Conf. Humanoid Robots*, 2008.

[17] L. Schillingmann, P. Wagner, C. Munier, B. Wrede, and K. Rohlfing, "Using prominence detection to generate acoustic feedback in tutoring scenarios," in *INTERSPEECH*, Firenze, Italy, 2011, ISCA.

[18] P. Mermelstein, "Automatic segmentation of speech into syllabic units," *J. Acoust. Soc. Am*, vol. 58, no. 4, pp. 880–883, 1975.

[19] H. Sakoe and S. Chiba, "Dynamic programming algorithm optimization for spoken word recognition," *IEEE Trans. Acoustics, Speech and Signal Processing*, vol. 26, no. 1, pp. 43–49, 1978.

[20] D. Talkin, "A robust algorithm for pitch tracking (RAPT)," *Speech coding and synthesis*, vol. 495, pp. 518, 1995.

[21] P. Boersma and D. Weenink, "Praat: doing phonetics by computer (v. 5.1.21) [computer program].," November 2009.