# Intonation Based Sentence Modality Classifier for Czech Using Artificial Neural Network

Jan Bartošek and Václav Hanžl

Department of Circuit Theory, FEE CTU in Prague, Czech Republic,
Technická 2, 166 27 Praha 6 - Dejvice, Czech Republic
{bartoj11,hanzl}@fel.cvut.cz
http://obvody.feld.cvut.cz/

**Abstract.** This paper presents an idea and first results of sentence modality classifier for Czech based purely on intonational information. This is in contrast with other studies which usually use more features (including lexical features) for this type of classification. As the sentence melody (intonation) is the most important feature, all the experiments were done on an annotated sample of Czech audiobooks library recorded by Czech leading actors. A non-linear model implemented by artificial neural network (ANN) was chosen for the classification. Two types of ANN are considered in this work in terms of temporal pattern classifications - classical multi-layer perceptron (MLP) network and Elman's network, results for MLP are presented. Pre-processing of temporal intonational patterns for use as ANN inputs is discussed. Results show that questions are very often misclassified as statements and exclamation marks are not detectable in current data set.

**Keywords:** sentence modality, intonation, temporal pattern classification, non-linear model, neural network.

## 1 Introduction

Prosodic information is still not sufficiently used in today's automatic speech recognition (ASR) systems. One possibility how to use prosodic information is to create the punctuation detection module. This work can be viewed as a basic feasibility study of prosodic "standalone" automatic punctuation detector for Czech language. "Standalone" property means that the module can be almost independent on hosting ASR system, because the punctuation detector will not use any of the information provided by ASR (recognized words and its boundaries, aligned phonemes durations, etc.) and will operate directly on raw acoustic data.

There are several studies dealing with punctuation detection. The first of these studies used only lexical information by building 3-gram language model [1] (and recently [2] with dynamic conditional random fields approach), others also utilized acoustic information [3], when acoustic baseforms for silence and breath were created and punctuation marks were then considered to be regular

C.M. Travieso-González, J.B. Alonso-Hernández (Eds.): NOLISP 2011, LNAI 7015, pp. 162–169, 2011.

words and added to the dictionary. [4] investigated that pitch change and pause duration is highly correlated with position of punctuation marks and that F0 is canonical for questions and used CART-style decision trees for prosodic features modelling. In [5] a detection of three basic punctuation marks was studied with combination of lexical and prosodic information. Punctuation was generated simultaneously with ASR output while the ASR hypothesis was re-scored based on prosodic features. Ends of words are considered as the best punctuation candidates. For this reason, all the prosodic features were computed near the word ends and in two time windows of length 0.2s before and after this point. The prosody model alone gives better results than the lexical one alone, but best results were achieved by their combination. Authors also mentioned complementarity of prosodic and lexical information for automatic punctuation task. Combination of prosodic and lexical features also appeared in [6] where punctuation process was seen as word based tagging task. Pitch features were extracted from a regression line over whole preceding word. Authors also mentioned evaluation metric issues and except for Precision and Recall (P&R, F-measure), they also used Slot Error Ratio (SER) as well. Language model in combination with prosody model reduces P&R and SER, especially with the pause model for full-stop detection. Maximum entropy model was presented for punctuation task as a natural way for incorporating both lexical and prosodic features in [7], but only pause durations were used as prosodic features. Lexical-based models performed much better then pause-based models which is in contrast to the other former studies. Work [8] presents approach for punctuation based only on prosody when utilizing only two most important prosodic attributes: F0 and energy. Method for interpolating and decomposing the fundamental frequency is suggested and detectors underlying Gaussian distribution classifiers were trained and tested. [9] continues in the idea and claims that interrogative sentences can be recognized by F0 (intonation) only and about 70% of declarative sentences can be recognized by F0 and energy.

A closely related task to the automatic punctuation is sentence boundary detection which is discussed in [10], where a pause duration model outperforms language model alone. Again, the best results are achieved by combining them.

The problem of detecting patterns in time series is also widely discussed and deals mainly with the time and amplitude variability of the observed patterns. There are studies that try to appropriately pre-process the time series in the scope of a sliding window and than run matching algorithm to compute distance from the searched patterns in defined metrics [11,12]. On the other hand, an artificial neural network (ANN) approach for finding patterns in time series was also developed in the past, especially by Elman [13] network type. [14] brings nice overview and introduction into problematic on either conversion the time domain into spatial one or utilization of memory (loopbacks) in network architectures. An example of the application of ANN approach could be [15] utilizing classical multi-layer perceptron and FIR based network or [16] dealing with financial stock time-series data.

In this article we use a slightly different approach than the other studies related to punctuation detection. Most of the previous works tend to benefit from knowledge of lexical information (words themselves and their time boundaries) mostly obtained from transcriptions or speech recognizer outputs. Then, even when using prosodic information, this information is word based (e.g. F0 range, slope within word). In contrast to this, we are trying to classify the modality of the sentence without knowledge of words and its boundaries as it was done in [9] and try to find out whether Czech intonation contours alone are sufficient cues.

## 2 Intonational Patterns

Sentence melody is language dependent. The intensity of Czech intonation varies according to the region and is also individual, but general trends across all these nuances are obvious. There is only a slight difference in definition of terms "melodeme" and "cadence" [17] in connection with intonational patterns. Cadence is an abstract scheme of a melody course and is created by the sequence of intonational changes, where the count and the direction of these changes are given. Melodeme is a term used for the basic type of intonational course connected with grammatical functions. In other words, melodeme is the set of melodic schemes that are used in language in the same type of sentences. The cadence is then used for one particular melodic scheme itself.

The cadence in a function of melodeme usually takes up only a part of an utterance. The place in the sentence marking the beginning of the cadence is the measure that can have sentence-type accent as the last one in the whole sentence. From this point the cadence drives the melodic course until the end of sentence is reached (for determining cadence). The length of the cadence is thus variable and a melodic course is distributed in relation to the syllable count of the cadence.

According to [17] there are three basic types of melodemes in Czech: concluding descending, concluding ascending and non-concluding. Each of them has various cadences and it is beyond the scope of this paper to go into details of linguistic theory. But the conclusion is that the set of melodemes unfortunately does not uniquely match the set of punctuation marks. For example, there are two types of questions with different melodemes, but single punctuation mark ('?'). Besides, there is one melodeme (concluding descending) standing for various modality types. This fact makes the task of sentence modality classification even more difficult. Also finding the beginning of the cadence could be a problematic task for speech processing.

## 3 Neural Networks for Temporal Processing

Artificial neural networks (ANN) are a well known tool for classification of static patterns, but could also be a good model for dealing with the time series. From theory ANN could be seen as non-linear statistical models. MLP networks can be considered as a non-linear auto-regressive (AR) model and can approximate

arbitrary function with arbitrary precision depending only on the number of units in the hidden layer. By training the network we are trying to find the optimal AR-model parameters.

Two basic approaches for the classification of temporal patterns are: 1) a usage of the classical MLP feedforward network or 2) usage of special type of neural network with 'memory'. In the first case we are dealing with a fixed number of inputs in the input layer of the network, where no 'memory' is available. This means we need to map time dimension onto spatial one by putting the whole fixed-length frame of signal onto all the inputs of MLP network. The main issue is that time patterns vary not only in amplitude, but also in its duration and thus we need to choose suitable frame length. In the next step another frame (depending on the shift of frames) of signal is brought on the inputs and the network gives a new answer with no connection to the previous one. In the recurrent types of network, there are loopbacks creating the memory. This architecture allows us to have only one input and bring one sample on it in each step and get new output.

## 4 Training and Testing Data

Although there are many databases for the training of ASRs, not many of them can be used for our task. Firstly, in most of the cases punctuation marks are missing in the transcriptions in these databases. This flaw can be removed by re-annotating the data and putting punctuation marks back in the right places. Secondly, there is often a shortage of prosody and modality rich material in these databases. And what is worse, if the material exists, the speakers in most cases do not perform the prosody naturally, because of the stress when being recorded. Special emotive databases exist too, where certain parts of it can be used, but emotions of speaker are not the object of our study. That is why alternative data sources were looked for.

Finally, the online library of Czech audiobooks read by leading Czech actors was used. A compressed MP3 format of audio files did not seem to be an obstacle as the records are very clean with studio ambient. In addition, actor's speech is a guarantee of intonation rich material. For first experiments presented in this paper a basic sample of the library including unified data from 4 different audiobooks read by 4 different actors (3 men, 1 woman) was manually annotated to roughly include a natural ratio of punctuation marks for Czech language. The counts of individual punctuation marks can be found in the table 1. As in the future we plan to increase the amount of data with use of an automatic alignment system based on available electronic versions of the books, we did not manually mark the places where beginning of the cadence occurs. This task would even need phonetic specialists assistance and it is very difficult to automate. That is why intonation pattern for corresponding following punctuation mark is taken from the beginning of the whole sentence or previous non-concluding punctuation mark (comma). Basic intonation contour was computed directly using PRAAT [18] cross-correlation PDA with default settings.

**Table 1.** Occurences of punctuation marks in the used data set

| **Punctuation mark** | **?** | **!** | **,** | **.** | sum |
|---|---|---|---|---|---|
| count | 31 | 7 | 65 | 158 | 261 |

## 5 Pre-processing the Intonation Patterns

Raw data pre-processing is a common first step to meet requirements of the task. When using the neural networks for pattern classification, there is also need to prepare the data to maximally fit the chosen network architecture.

1. **Logarithmic scale conversion**
   Due to the fact that a human perception of pitch occurs in roughly logarithmic scale, we need to convert frequency values (in Hz) into musical scale values (semitone cents) according to equation 1, where ideally $f_{LOW}$ is a low frequency border of vocal range of the speaker. This makes the signal values relative to this threshold and deletes differences of absolute voice heights (curves of same patterns should now look the same even for man or women speech). This conversion also implicitly removes the DC component of intonation signal, but it also means we need to know what the lowest frequency border of the vocal range of the speaker is. From training and testing data sets this can be computed as finding minima over all of the units spoken by the speaker. When applied online, we will gradually make the estimate of this value more and more accurate.

$$Cents = 1200 \log_2(\frac{f}{f_{LOW}}) \tag{1}$$

2. **Trimming the edges**
   As the annotated patterns have silent passages on the beginning and at the end (zero-valued non-voiced frames), we need to remove these parts of the signal for further processing (see the next step).
3. **Interpolating missing values**
   The speech signal consists not only of voiced frames when the glottis do pulse with certain period, but also of unvoiced frames when the glottis do not move (unvoiced consonants). Good pitch detection algorithm can distinguish between these two cases. This leads to situation of having zero values as a part of the intonation curve. Such a curve does not seem to be continuous even for very fine time resolution. Because these "zero moments" depend on certain word order and not on supra-segmental level of sentence, we need to get rid of them and thus maintain that same intonation pattern with another words in it leads to the same final continuous intonation pattern.
4. **Removing micro-segmental differences of intonation**
   As we are following intonation as supra-segmental feature of speech, we are not interested in intonation changes that occur on intra-syllable level. That is why we want to erase these fine nuances and maintain only the main

character of the curve. This can be accomplished by applying an averaging filter on the signal. We could also achieve similar result by choosing longer signal window and its shift in pitch detection algorithm setting.

5. **Reconstructing the levels of extremes**
   Previous smoothing unfortunately also smoothed out the intonation extremes, changing their original pitch. Because these extremes are very important for pattern character, we want to 'repair' them. In current implementation only two global extremes are gained to their former values by adding (subtracting) appropriately transformed Gaussian curves with height of differences between original and smoothed values and with width of previously used smoothing filter.
6. **Signal down-sampling**
   High time resolution of time patterns leads to a need of a high number of inputs for classical MLP or long 'memory' for recurrent ANN. Both facts imply a higher unit count in both types of network, which is then more difficult to train with limited amount of training data. That is why down-sampling of pattern is needed. Down-sampling is done several times according to the type and architecture of ANN used for follow-up classification:
   (a) 'Normalizing' down-sampling - MLP type of network with temporal into spatial domain conversion needs fixed length vector on its input. Each pattern is thus normalized in its length to satisfy the 64 or 32 input vector length condition.
   (b) Classical down-sampling - recurrent networks do not require fixed-length patterns, but to perform reasonably, too precise time resolution of the series implies high count of hidden units . That is why a classical down-sampling from 1000Hz sampling rate to 40Hz and 25Hz is done.

## 6 Results and Discussion

The experiments were made on the data set, where 70% of it were training data, 15% validation set and 15% test data. Trained network was then evaluated on the whole data set. Results for intonational patterns with fixed length of 32 samples on MLP with 15 units in hidden layer can be seen from the confusion matrix (table 2) evaluated over the whole data set. The classifier tends to prefer classes with higher occurrence in training data set (commas) due to their statistically higher occurrence in validation set. That is why an another experiment was done using a limited equal distribution of the patterns in the classes (all the classes contain 31 patterns except the exclamation mark class). Representative confusion matrix for the reduced data set is in the table 3 for 32 samples per pattern and 20 hidden units. From the results it is obvious that the MLP network is capable to give near a 50% success classification rate for classes of question marks, commas and full stops. The impossibility of classifying exclamation marks could be based on the fact that these intonation patterns are not stable in intonation and that this type of modality rather lies in another prosodic feature (energy), or the data set for this class was too small in our corpus. The last experiment was

**Table 2.** MLP Confusion matrix in %, full data set, full pattern length

| **Actual class →** **Predicted class ↓** | **?** | **!** | **,** | **.** |
|---|---|---|---|---|
| **?** | **10** | 0 | 2 | 1 |
| **!** | 0 | **0** | 0 | 0 |
| **,** | 20 | 29 | **18** | 6 |
| **.** | 70 | 71 | 80 | **93** |

**Table 3.** MLP Confusion matrix in %, reduced data set, full pattern length

| **Actual class →** **Predicted class ↓** | **?** | **!** | **,** | **.** |
|---|---|---|---|---|
| **?** | **40** | 32 | 32 | 24 |
| **!** | 1 | **4** | 1 | 1 |
| **,** | 33 | 41 | **48** | 23 |
| **.** | 26 | 23 | 19 | **52** |

**Table 4.** MLP Confusion matrix in %, reduced data set, last 1200ms of pattern

| **Actual class →** **Predicted class ↓** | **?** | **!** | **,** | **.** |
|---|---|---|---|---|
| **?** | **46** | 30 | 26 | 28 |
| **!** | 1 | **7** | 1 | 1 |
| **,** | 28 | 40 | **57** | 22 |
| **.** | 25 | 23 | 16 | **49** |

done on cut-length patterns, where only last N={1500,1200,800,500}ms were left, then down-sampled to 32 and 64 samples for MLP input. 64-sample patterns were more successfully recognized. Best results for the reduced data set were obtained for 1200ms patterns and 10 hidden neurons (table 4).

## 7 Conclusion and Future Work

We discussed two approaches for the classification of sentence modality based purely on the intonation. MLP based approach gives classification success rate around 50% on question mark, comma and full stop classes. As expected, questions are often misclassified as statements. Results show that there is need to think about possible improvements. This could be probably done in various ways: larger set of training data, better pre-processing of intonation contour, different ANN architecture (step-by-step Elman network approach) or joining another prosodic features besides intonation alone (energy, pause duration). After getting more satisfying results we would also like to include the model to the block for online punctuation detection working next to speech recogniser.

**Acknowledgments.** Research described in the paper was supported by the Czech Grant Agency under grant No. 102/08/0707 Speech Recognition under Real-World Conditions and grant No. 102/08/H008 Analysis and modelling of biomedical and speech signals.

## References


1. Beeferman, D., Berger, A., Lafferty, J.: Cyberpunc: A lightweight punctuation annotation system for speech. In: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, pp. 689–692 (1998)
2. Lu, W., Ng, H.T.: Better punctuation prediction with dynamic conditional random fields. In: Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, EMNLP 2010, pp. 177–186. Association for Computational Linguistics, Stroudsburg (2010)
3. Chen, C.J.: Speech Recognition with Automatic Punctuation. In: Proc. Proc. 6th European Conference on Speech Communication and Technology, EUROSPEECH 1999, Budapest, Hungary, pp. 447–450 (1999)
4. Shriberg, E., Bates, R., Stolcke, A., Taylor, P., Erringer, A., Gregory, M., Heintzelman, L., Metzler, T., Oduro, A., The, T.: Can prosody aid the automatic classification of dialog acts in conversational speech? Language and Speech 41(3-4), 439–487 (1998)
5. Hwan Kim, J., Woodland, P.C.: The use of prosody in a combined system for punctuation generation and speech recognition. In: Proc. EUROSPEECH 2001, pp. 2757–2760 (2001)
6. Christensen, H., Gotoh, Y., Renals, S.: Punctuation annotation using statistical prosody models. In: Proc. ISCA Workshop on Prosody in Speech Recognition and Understanding, Red Bank, NJ, USA, pp. 35–40 (2001)
7. Huang, J., Zweig, G.: Maximum Entropy Model for Punctuation Annotation from Speech. In: Proc. International Conference on Spoken Language Processing (ICSLP 2002), pp. 917–920 (2002)
8. Strom, V.: Detection of accents, phrase boundaries and sentence modality in german with prosodic features. In: EUROSPEECH, vol. 3, pp. 3029–2042 (1995)
9. Král, P., Cerisara, C.: Sentence modality recognition in french based on prosody. In: VI International Conference on Enformatika, Systems Sciences and Engineering, ESSE 2005, vol. 8, pp. 185–188. International Academy of Sciences (2005)
10. Gotoh, Y., Renals, S.: Sentence boundary detection in broadcast speech transcripts. In: Proc. of ISCA Workshop: Automatic Speech Recognition: Challenges for the new Millennium ASR 2000, pp. 228–235. International Speech Communication Association (2000)
11. Harada, L.: Complex temporal patterns detection over continuous data streams. In: Manolopoulos, Y., Návrat, P. (eds.) ADBIS 2002. LNCS, vol. 2435, pp. 401–414. Springer, Heidelberg (2002)
12. Jiang, T., Feng, Y., Zhang, B.: Online detecting and predicting special patterns over financial data streams. Journal of Universal Computer Science - J. UCS 15(13), 2566–2585 (2009)
13. Elman, J.L.: Finding structure in time. Cognitive Science 14(2), 179–211 (1990)
14. Dorffner, G.: Neural networks for time series processing. Neural Network World 6, 447–468 (1996)
15. Haselsteiner, E., Pfurtscheller, G.: Using time-dependent neural networks for EEG classification. IEEE Transactions on Rehabilitation Engineering 8(4), 457–463 (2000)
16. Zhou, B., Hu, J.: A dynamic pattern recognition approach based on neural network for stock time-series. In: NaBIC, pp. 1552–1555 (2009)
17. Palková, Z.: Fonetika a fonologie češtiny, Karolinum, Praha (1994)
18. Boersma, P.: Praat, a system for doing phonetics by computer. Glot International 5(9/10), 341–345 (2001)