

Automatic sentence modality recognition in children's speech, and it's usage potential in the speech therapy

Dávid Sztahó¹, Katalin Nagy¹, Klára Vicsi¹

¹ Laboratory of Speech Acoustics, Budapest University of Technology and Economics,
Department of Telecommunications and Media Informatics, Stoczek u. 2, 1111 Budapest,
Hungary
sztaho@tmit.bme.hu, menjus@gmail.com, vicsi@tmit.bme.hu

Abstract. In the Laboratory of Speech Acoustics prosody recognition experiments have been prepared, in which, among the others, we were searching for the possibilities of the recognition of sentence modalities. Due to our promising results in the sentence modality recognition, we adopted the method for children modality recognition, and looked for the possibility, how it can be used as an automatic feedback in an audio - visual pronunciation teaching and training system. Our goal was to develop a sentence intonation teaching and training system for speech handicapped children, helping them to learn the correct prosodic pronunciation of sentences. In the experiment basic sentence modality models have been developed and used. For the training of these models, we have recorded a speech prosody database with correctly speaking children, processed and segmented according to the types of modalities. At the recording of this database, 59 children read a text of one word sentences, simple and complex sentences. HMM models of modality types were built by training the recognizer with this correctly speaking children database. The result of the children sentence modality recognition was not adequate enough for the purpose of automatic feedback in case of pronunciation training. Thus another way of classification was prepared. This time the recordings of the children were sorted rigorously by the type of the intonation curves of sentences, which were different in many cases from the sentence modality classes. With the new classes, further tests were carried out. The trained HMM models were used, not for the recognition of the modality of sentences, but checking the correctness of the intonation of sentences pronounced by speech handicapped children. Therefore, an initial database, consisting of the recordings of the voices of two speech handicapped children had been prepared, similar to the database of healthy children.

Keywords: Speech Prosody Recognition, Automatic Speech Recognition, Prosody Database, Speech Technology, Hidden Markov Models

1 Introduction

The latest results in computer technology and digital speech processing made possible the construction of computer aided systems that improve the quality of speech learning and training. Speech training using this speech technology is useful for

speech handicapped persons, especially hard of hearing and deaf children. These children partially or fully cannot learn the way of correct pronunciation, including not only the phonemes, but also larger segmental features, like prosody because of the distortion of auditory feedback. Recently, speech technology has appeared in second language learning too. The capabilities of the technology include measuring and displaying the dynamic characteristics of speech parameters, using auditive, visual and automatic feedback [1]. Experiments have shown, that for example a visual F0 display of supra-segmental features combined with audio feedback is more effective than audio feedback alone [2][3], especially if student's F0 contour is displayed along with a reference model. The feasibility of this type of visual feedback has been demonstrated by a number of simple prototypes [14] [15].

Our main goal is to develop a speech teaching system for speech handicapped children through a computer-aided system. This way, we are going to teach the correct pronunciation of the different sentence modalities by visual and automatic feedback. Institutions that treat speech handicapped children, for example with hard of hearing disability, will be able to use the method for improving the children's pronunciation. To determine the final use of the method, like frequency of use and exact usage, needs discussion with experts on speech therapy of children.

Until now, teaching systems that have automatic feedback are based only on segmental features, not on supra-segmental features. One group of these systems takes advantage of the acoustic similarity between the trainee's acoustic production and a template to measure the correctness of trainee's production. In SPECO [4] and ISTR [5] similar metric was estimated. In the other group of speech training systems, phoneme-based Hidden Markov Models are applied [16] [17].

In this article we looked for the possibility, how an automatic prosodic recognizer, actually a sentence modality recognizer can be used for the automatic feedback in a pronunciation teaching system.

In the first part of our research, a prosodic database of children was prepared. It consisted of the recordings of healthy children. The recorded database was segmented and labeled manually according to the Hungarian sentence modality classes.

After the database preparation a sentence modality recognition test was carried out by an automatic prosodic recognizer, developed earlier in our Laboratory [6]

The result of the children sentence modality recognition was not adequate enough for the purpose of automatic feedback in case of pronunciation training.

Thus another way of classification was prepared. This time the recordings of the children were sorted rigorously by the type of the intonation curves of sentences, which were different in many cases from the sentence modality classes. With the new classes, further tests were carried out, by training and testing the prosodic recognizer. This way the recognition correctness increased to a great extent.

In order to test the real usage of the recognizer, a small database was made, consisting of recordings of two speech impaired children. The structure of this database corresponded with the first database. A method was examined with which the recognizer can be used for teaching prosody.

2 Former recognition system

For sentence modality recognition a formerly developed automatic prosodic classifier was used [6]. This recognition system is based on prosodic Hidden Markov Models. The features used for the training of these prosodic models, and for the recognition are fundamental frequency and energy values, their derivatives and second derivatives. The prosodic HMM models were built using HTK toolkit [8].

To train this recognizer, speech databases were processed according to the types of modalities that were used. An HMM was assigned to each modality, then the aligned modality sequence of the incoming speech was searched for.

For the extraction of fundamental frequency and energy values the Snack toolkit was used [7]. The calculation was done with 150 ms window size and 10 ms time step. Fundamental frequency values were corrected by anti-octave filtering and median filtering. The recognition was done for six sentence modality classes – declarative, question to complement, yes-no question, imperative, sentence with wish (“if only...”) and clause (not closing) – and one silence model. The recognizer was trained with MRBA [9] and BABEL [10] databases.

3 Database

It is necessary to have a correctly built database, for every statistical-based recognition system, which consists of the acoustic signals needed to train and test the recognition system. Because our goal was to develop a sentence modality recognizer, we needed a database consisting of recordings of the appropriate sentence modality classes.

Two text materials were prepared. The first text contained on average 10 individual sentences from each modality class. These sentences were obtained from SPECO [11], software developed in our laboratory for children speech therapy. The second text contained 3 short dialogues, made from the sentences of the first text. All of the sentences were used. Some more sentences were needed to be added in order to build appropriate dialogues. These added sentences were not used during the recognition experiments.

The recording condition is shown on figure 1. Children were sitting in front of a MONACOR ECM-100 microphone and read both texts. Every sentence was recorded twice, in order to ensure at least one good recording from each sentence. The microphone was connected to a CREATIVE Sound Blaster Audigy 2 NX external sound card which was further connected to a portable computer.

The database was recorded in an elementary school at Budapest [Elementary School of Farkasrét in Budapest] with 60 healthy and correctly speaking children. The age of the children was between 8 and 13. This database was used for training the recognition system and evaluating it. For further use, the database was manually annotated by marking the sentence modalities for each recording. After annotating all sentences, the average number of samples per each class was 1600.

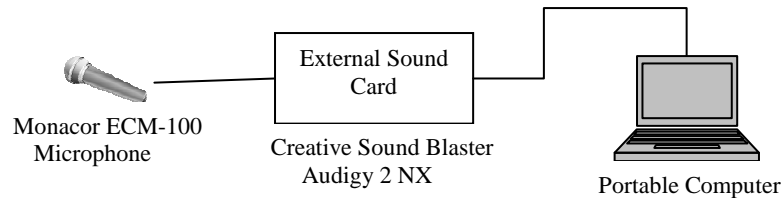


Fig 1. Recording condition

4 Adaptation

At the evaluation phase, 70 % of the samples were used for the training of the recognizer, and 30% was used for testing. The results are shown on table 1.

Table 1. Sentence modality recognition results. Classes: S: Declarative K: Question to complemented E: Yes-no question FF: Imperative O: Sentence with wish („if only...”) T: Clause (not closing) U: Silence

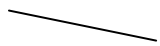
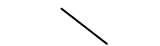
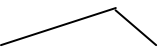
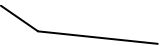
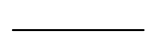
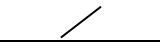
	S	K	E	FF	O	T	U	Corr (%)
S	895	279	91	353	240	41	6	47
K	49	579	50	103	68	36	1	65,3
E	65	158	862	95	110	34	2	65
FF	69	98	52	505	126	44	4	56,2
O	25	12	6	4	131	7	0	70,8
T	9	35	14	21	51	255	12	64,2
U	106	56	45	51	42	46	5791	94,4

Without the quiet parts, the overall result was 61.42 percent. This performance is good for a system which is for example an added system to a speech recognizer, but it is not acceptable for a pronunciation teaching system. Two major difficulties occurred during the evaluation tests. First, in some cases there were differences in intonation contours between samples belonging to the same class, because not all the children said the same modality class with the same prosodic characteristics. Second, the complexity of the sentences was different, there were one-word, simple and complex sentences in the case of some classes.

In order to solve these problems, new classes were created, replacing the former ones. The new classes were based strictly on the intonation contour of the sentences.

The new classes are shown in table 2. Examples of fundamental frequency contours can be seen on figures 2 and 3.

Table 2. New classes

TYPE OF INTONATION	TYPE OF SENTENCE	EXAMPLE	FORM OF INTONATION
Descending	declarative sentence	“Anikó is standing at the gate”	
Falling	question to be complemented	“Why is Anikó standing there?”	
Ascending-falling	yes-no questions	“Is Anikó standing at the gate?”	
Falling-descending	imperative and exclamation sentences	“Come here quickly!”	
Floating	clauses (not closing)	“Anna is standing at the corridor, ...”	
Rising	one word questions	“No?”	

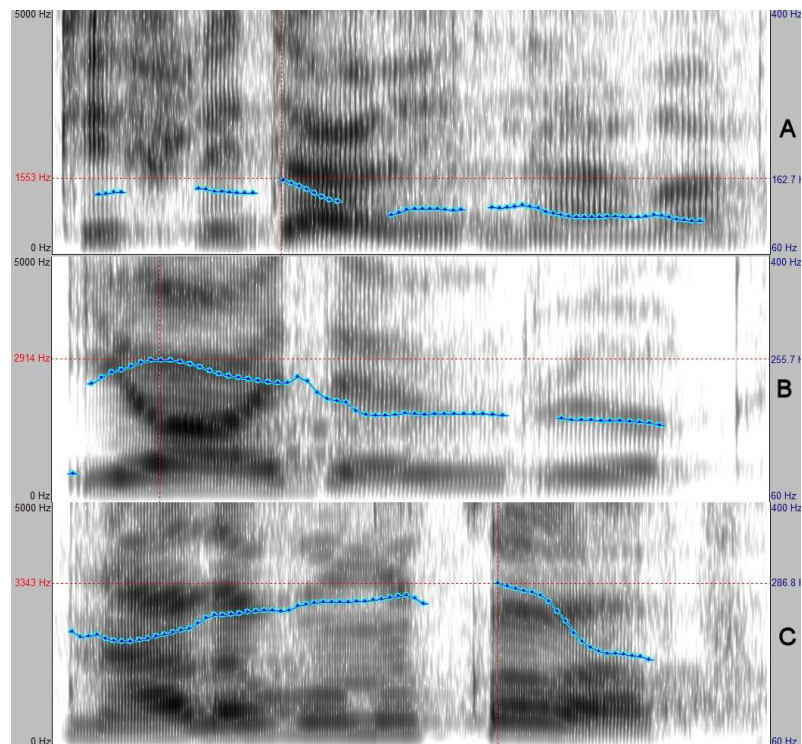


Fig 2. Examples for classes of table 2. (A: descending; B: falling; C: ascending-falling)

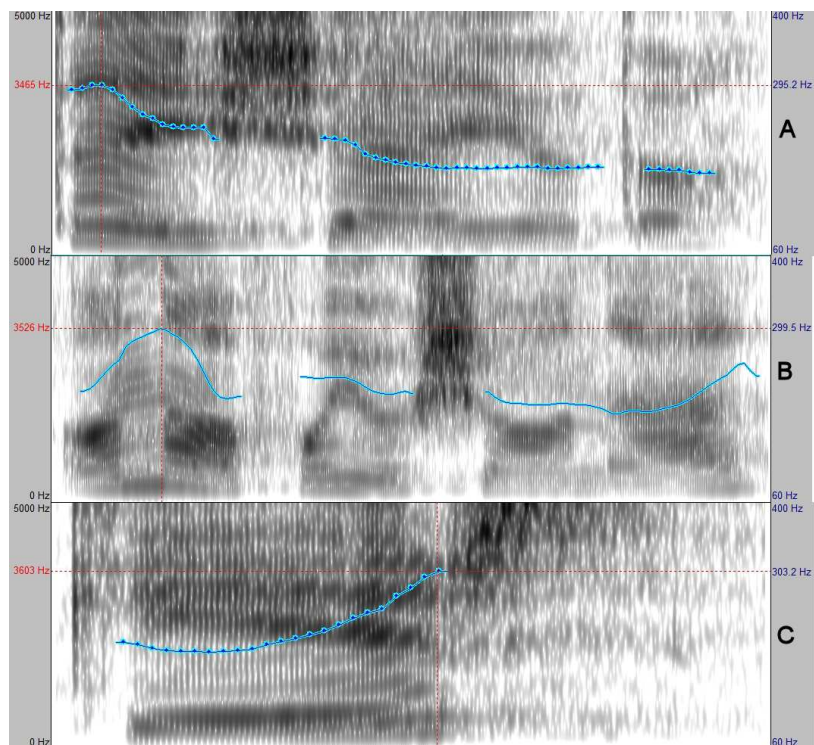


Fig 3. Examples for classes of Table 2. (A: falling-descending; B: floating; C: rising)

For the new recognition tests the sentences were selected as follows. From each class there was 1 given sentence selected per child. On average, each child said each sentence two times. If a given child did not read a sentence correctly, another sentence was selected from the same child. Because there were 60 children in the database, there were on average 120 sentences selected per class. As a result of this, we got a sentence set, that hereafter we will call “sentence set 1”. The above process was repeated on different sentences per class, and this way we got another sentence set, that hereafter we will call “sentence set 2”. With these two different sentence sets it was possible to tests grammatically similar and diverse sentences by mixing them together. The two sentence sets are shown in Table 3.

Table 3. Sentences of the two sentence sets

Class	Sentence set 1	Sentence set 2
A	David has the croissant.	Peti has got the hat.
B	Oh, how am I glad!	If only he was standing there!
C	Does he have got the hat?	Did you get a mark today?
D	You must stand there too!	Come here quickly!
E	Anna is standing at the corridor, (...)	Anna goes with you, (...)
F	Want some?	But?

Table 4 shows, that during the initial tests, in case of some classes, the automatic recognizer could not separate classes correctly, in the case of classes that have similar prosodic characteristics. These classes were the “Falling” and the “Falling-descending”. Further on these classes were contracted.

Table 4. Recognition results with sentence set 1 for six classes. Dark fields show, that in case of classes B and D the system could not decide unequivocally. (Classes: A: descending; B: falling; C: ascending-falling; D: falling-descending; E: floating; F: rising)

	A	B	C	D	E	F	CORR [%]
A	40	0	1	1	1	0	93.0
B	0	25	0	7	0	0	78.1
C	0	0	28	1	0	0	96.6
D	1	15	0	13	1	0	43.3
E	5	2	1	1	30	0	76.9
F	0	0	0	0	0	36	100.0

With the reduced class set (B and D are contracted) the recognition results were above 90 percent (table 5). When the two sentence sets were mixed, the recognition went below 87 percent. In order to see the effect of more diverse training data, a third sentence set was chosen in the same way as the other two. Figure 4 shows, what kind of effect has the more diverse training data on the recognition. It shows that the best recognition can be achieved by using less diverse training and testing data. Therefore in the real teaching system, only a specific data set should be used at a time.

Table 5. Recognition results with sentence set 1 for reduced class set

	A	B	C	E	F	CORR [%]
A	42	1	0	0	0	97.7
B	1	51	5	4	1	82.3
C	0	0	27	1	1	93.1
E	1	1	1	36	0	92.3
F	0	0	0	0	36	100

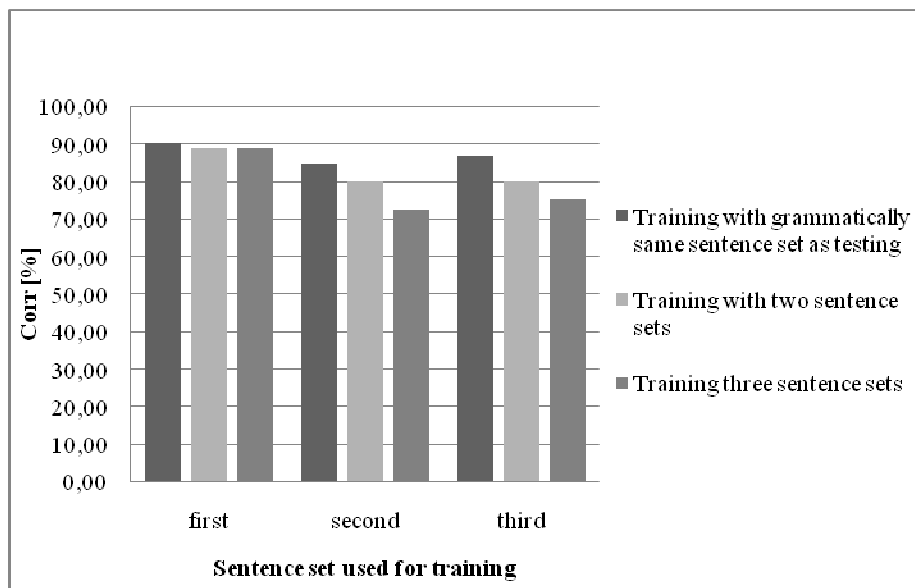


Fig 4. Effect of the diverse training data on automatic recognition

5 Testing method with hard of hearing children

The main goal of the research is to develop a recognition system that can be used in a teaching system for hard of hearing children. In order to do this, the recognition system built in the previous chapter was used for a small database of speech handicapped children's recordings, which was recorded in a special institute for hearing handicapped children. This database consisted of recordings from two speech impaired children. The correct operation of the recognizer means that if the child says the sentence with bad pronunciation, the recognizer will classify it falsely. On the other hand, if the pronunciation is good, the recognizer has to classify it correctly. On Figure 5, there is an example shown, how the recognition system can be used in a real environment. The first fundamental frequency contour is from a healthy child, the second is from a hard of hearing child, and both of them were classified correctly. The third recording is from a hard of hearing child, and the recognizer did not classify it correctly to the same class as the others.

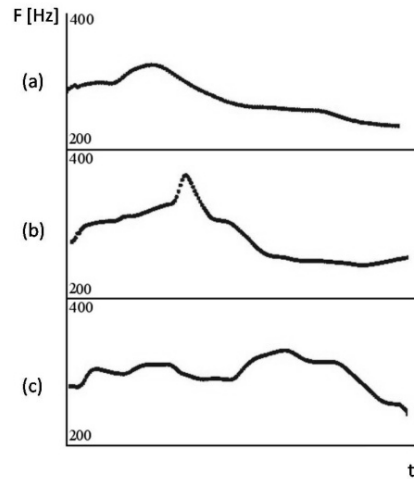


Fig 5. Recognition of recording of hard of hearing children illustrated by fundamental frequency. (Description: (a) recording from healthy child; (b) a correctly classified recording from hard of hearing child; (c) a falsely classified recording from hard of hearing child)

It can be seen from the examples presented above that the problem of the automatic feedback is that we do not know when the recognizer has to refuse the pronunciation, and when the recognizer can accept the actual one. It must be worked out in the future. Otherwise the strictness can be adjusted by modifying the probability of the acceptance.

6 Conclusion

In the paper we presented an automatic sentence intonation recognition method, that can be used as an automatic feedback in a prosody teaching system in the future. First, a database was created that consisted of recordings from healthy children. The text material of the database was assembled from two texts. The first consisted sentences of six Hungarian modality classes. The sentences were selected from formerly developed software, SPECO. The second text was made from the first text, making three dialogues using the same sentences. The built database was segmented and annotated marking the sentence modality classes.

After the initial tests it was clear that the original six sentences modality classes used for speech recognition purposes were not appropriate, therefore a new annotation was made. Sentences were categorized strictly according to their intonation, and thus the recognition increased to a great extent.

Verifying tests were made with the database of the healthy children in order to prepare the recognizer to its real usage. The results showed that the recognition ratio decreased in the case of more diverse training data. Thus it is clear, that for pronunciation teaching purposes only well selected training data are acceptable for the contraction of the reference sentence intonation models in the recognizer.

In the future, it is necessary to develop an optimization technique on the base of which we can adjust the probability of the acceptance of the recognizer. Thus we plan to prepare a large sentence intonation database of impaired children. Moreover for tuning the probability of the acceptance of the automatic recognizer we plan a perception experiment, in which human listeners will have to categorize the different sentences by listening to them. This perception experiment makes us able to examine the listener's acceptance of different sentences.

Acknowledgement

We would like to thank to "Dr. Béla Török" Kindergarten, Elementary School and Special School, and Elementary School of Farkasrét making available for us to make recordings with the children.

References

1. Vicsi, K. Computer-Assisted Pronunciation Teaching and Training Methods Based on the Dynamic Spectro-Temporal Characteristics of Speech. *Dynamics of Speech Production and Perception*, IOS Press. pp. 283-304. 2006.
2. de Bot, K. Visual feedback of intonation: Effectiveness and induced practice behavior. *Lang. Speech* 26(4):331-335,1983.
3. James, E. The acquisition of prosodic features of speech using a speech visualizer. *IRAL*, 14(3):227-243, 1976.
4. Vicsi, K., Csatári, F., Bakcsi, Z. and Tantos, A. Distance score evaluation of the visualized speech spectra at audio-visual articulation training. *Proc. Eurospeech*, pp. 1911-1914, 1999.
5. ISTR Indiana Speech Training Aid Features. Bloomington, IN: Communication Disorders Technology, Inc. http://www.comdistec.com/istra_faq.shtml, 2003.
6. Vicsi, K.; Szaszák, Gy. Using Prosody for the Improvement of ASR - Sentence Modality Recognition, *Proc. of Interspeech2008*, Bristol, 2008, -ISCA Archive, <http://www.isca-speech.org/archive>
7. The Snack Sound Toolkit, <http://www.speech.kth.se/snack/>
8. HTK Speech Recognition Toolkit, <http://htk.eng.cam.ac.uk/>
9. Vicsi K., Velkei Sz., Szaszák Gy., Borostyán, G., Gordos G. Speech recognizer for preparing medical reports. Development experiences of a Hungarian speaker independent continuous speech recognizer. *Híradástechnika*, 2006/7, (p. 22-27), 2006
Hungarian Reference Speech Database (MRBA)
<http://alpha.tmit.bme.hu/speech/hdbMRBA.php>
10. Roach, P., Vicsi, K., Gordos, G. Report on BABEL, an Eastern European Multi-language database. COST 249 meeting, Zurich, 17-18 October, 1996.
11. Vicsi K., Váry Á. Distinctive training methods and Evaluation of a Multilingual, Multimodal Speech Training System, *SPECO Proc. 4th Intl. Conf. Disability, Virtual Reality and Assoc. Tech.*, Veszprém, Hungary, (p. pp. 47.-52.), 2002.
12. Szaszák, Gy.; Vicsi, K. Speech recognition supported by prosodic information for fixed stress languages, *proceeding of TSD conference Brno*, (p. 262-269), 2007.
13. Szaszák, Gy.; Vicsi, K. Using Prosody in Fixed Stress Languages for Improvement of Speech Recognition, *Workshop Vietri*, 2007.

14. Anderson-Hsieh, J. Interpreting visual feedback in computer-assisted instruction on suprasegmentals. *CALICO Journal*, 11(4):5-22, 1994.
15. Hiller, S., Rooney, E., Lefevre, J.P. and Jack, M. SPELL: An automated system for computer-aided pronunciation teaching. *Proc. Eurospeech*, 1993.
16. Kawai, G. and Hirose, K.A. CALL system using speech recognition to train the pronunciation of Japanese long vowels, the mora nasal and mora obstruents. *Proc. Eurospeech*, 1997.
17. Narusa, J. Computer-aided spoken language training with enhanced visual and auditory feedback. *Proc. Eurospeech*, pp. 183-186, 1999.