interspeech 2008 incorporating SST 08

USING PROSODY FOR THE IMPROVEMENT OF ASR – Sentence Modality Recognition

Klára Vicsi, György Szaszák

Laboratory of Speech Acoustics, Department of Telecommunications and Media Informatics, Budapest University of Technology and Economics

vicsi@tmit.bme.hu, szaszak@tmit.bme.hu

10.21437/Interspeech.2008-136

Abstract

In the Laboratory of Speech Acoustics ASR research has been prepared, in which we were searching for the possibility to contribute to the higher linguistic processing levels of ASR – at syntactic, and semantic level – by acoustical pre-processing of the supra-segmental (prosodic) features. The subject of our current article is a semantic level processing, built on supra-segmental parameters. HMM models of modality types of sentences were built by training the recognizer with speech databases processed according to the types of modality, and a simple set of connection rules of modalities were used as linguistic model. The best recognition results were obtained, when state numbers of HMM clause type-models were 11, and each state had 2 Gaussian components. With these adjustments the accuracy of recognized types of modalities was 71 % for Hungarian, and 78% for German, even though the database was small for both languages.

Index Terms: speech recognition, prosody, sentence modality, HMM models

1. Introduction

There are numerous levels of the speech recognition process: acoustic, phonetic-phonological, syntactic, semantic, and pragmatic level [1]. The more we can involve from these levels into the automatic speech recognition process, the more accurate the recognition will be.

In the automatic speech recognition the acoustic pre-processing unit works on acoustic level, and performs the analysis of speech signal, the compressing and the feature extraction. The extracted parameters appear on its output (feature vectors) by time frames. These parameters are spectral parameters, most usually MFC (Mel Frequency Cepstral) coefficients over a 25-50 ms time window, and typically measured at each 10 ms time frame in case of a segmental acoustic level pre-processing (cf. "a" processing limb of Figure 1.) [2], [3].

At the next level at the phonetic-phonological level, we make the creation of models of phonemes with the help of the pre-processed parameters obtained at the acoustic level. We compare the feature vector sequences obtained at the acoustic level with models of phonemes. This is the processing level, where we can get a phoneme sequence at the output. If we connect a syntactic level grammar to the recognition – for example, the N-gram language models used the most widespread, – then we will get a word sequence at the output [4], as it can be seen in the "a" limb of Figure 1. This is the way of the operation of speech recognizers available at the today's trade.

In the Laboratory of Speech Acoustics, a research has been prepared, in which we were searching for the possibility to contribute to the higher linguistic processing levels – syntactic and semantic levels – by acoustical pre-processing of supra-segmental (prosodic) features.

Using prosodic features in automatic speech recognition is not a trivial task, however, several attempts proved to be successful in this domain. Veilleux and Ostendorf presented an N-best rescoring algorithm [6] based on prosodic features and they have significantly improved hypothesis ranking performance. Kompe et al. presented a similar work for German language in [7]. Gallwitz et al. described a method for integrated word and prosodic phrase boundary recognition [8].

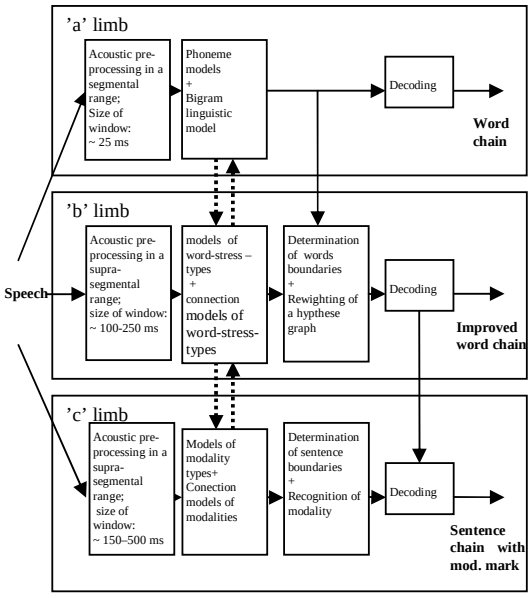

Figure 1. Block diagram of a multilevel extended speech recognizer

At syntactic and semantic processing levels, which are ranked higher in language hierarchy, it is not practical to use the feature vectors of the acoustical feature extraction made at segmental range at the acoustic-phonetic level mentioned

above. There is a need for other acoustical feature extraction based on supra-segmental (prosodic) features, which reflect the differences between the particular speech contents, the articulation by meaning, and even the emotions, too. Accordingly it is reasonable to measure the acoustical parameters of speech in the supra-segmental range (supra-segmental parameters), characteristically in more broad frequency- and time resolution, than in segmental range, when our purpose is the segmental characterisation of phoneme-like units. For that very reason a part of our research was to find the optimal time resolution for such a measurement of the acoustical parameters in the supra-segmental range.

When we use acoustical parameters of speech in the supra-segmental range in order to support linguistic processing in speech recognition, we can do it on several levels, as it is shown on Fig. 1 in limbs „b" and „c". Approaching the higher linguistic levels we can contribute step by step to making speech recognition more robust. At the syntactic level we can significantly improve the accuracy of the recognition, compared to the accuracy of a recognizer, which have a traditional word string output, by rescoring of word hypothesis graphs, in case of fixed stress languages (e.g. Hungarian or Finnish), when recognition algorithm is extended by automatic marking of word boundaries. This syntactic level processing can be followed up in „b" processing limb of Fig. 1. In several fixed stress languages, word beginnings can be marked automatically with reliable accuracy by word-stress detection, furthermore, by this, the marking of word boundaries and the rescoring of the word hypotheses graph obtained in the limb „a" is possible, which increases the robustness of the recognition. Details have been written in earlier reviews [9, 10].

In the „c" processing limb of Fig. 1 a semantic level processing is shown. The subject of our current article is just this semantic level processing based on supra-segmental parameters, and during this we could execute recognition of modality of sentences and localization of boundaries of the occurred sentences and clauses. During the recognition of modality we could determine, if word range had been said in a form of a statement, question, or an exclamation. By localization of sentence and clause boundaries we will get a guideline according to that which are the beginning and end times of clauses or sentences in the word chain.

2. Development of a semantic level modality recognizer

We have implemented sentence modality for Hungarian and for German on the basis of a statistical principle, using the Hidden Markov Model (HMM) method, for which we applied the HTK development system tool [3]. To train the recognizer, speech databases, processed according to the types of modality were used. The simple sentences were marked according to their modality. The complex sentences were divided into clauses. The final part of the complex sentence was marked according to the sentence modality. Clauses, before the closing ones and enumerations were formed a separate group. See Table 1.

HMM models of modality types were built by training the modality recognizer with the above mentioned databases. Simple set of connection rules of modalities were used as linguistic-like constraints additionally, describing the access possibilities of clauses and simple sentences.

Optimization of the parameters of the HMM modality models was considered as a basic task of our work.

2.1. Making of the training material

For Hungarian language, sentences of various modalities were collected from databases the BABEL [11] and MRBA [12], and for German language from the 1st volume of the KIEL Corpus [13]. 6 types of basic sentence modalities and clauses were distinguished for Hungarian and 4 for German according to Table 1. For German mostly "yes-no" questions were found and "Question to be complemented" were not present in a valuable number, and neutral modality was not differentiated.

Table 1. Data of segmentation and labelling

Modality of a simple sentence and a complex sentence by clause	**Mark-up (label)**	**Number of occurrences**	
		HUN	**GER**
Declarative sentence Closing clause of a declaration (int contour: descending HUN, GER)	S	445	404
Clauses, before the closing clause, enumerations (int contour:floating or floating-slow rising HUN, GER)	T	352	336
Closing clause of a question to be complemented(int contour: fall-descending HUN)	K	53	-
Yes-no question or closing clause of a yes-no question (int contour: rise-fall HUN, final rise GER)	E	35	89
Imperative and exclamation sentences, or closing clause of an imperative or exclamation sentence(int contour: rise-descending HUN, GER)	FF	52	63
Neutral	N	125	-
Total:		**1029**	**892**

The selected files were segmented and labelled by an expert on the basis of listening and of measuring of the fundamental frequency, or the energy level.

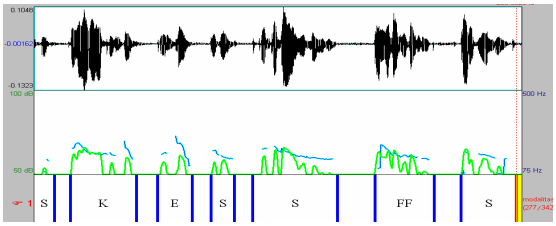

Figure 2: Segmentation and labelling in Praat program

The boundaries of sentences and clauses and the symbols appropriate to the types of modalities were marked in the speech material. Fig. 2 shows an example for the segmentation and labelling. In the first row of Fig. 2 we can see the waveform of the speech, and in the second row the diagram of the fundamental frequency and intensity. In the third row we can found the segmentation and labelling made by hand. We made the training of our prosodic recognizer with a database processed this way. Thus we generated 6 modalities and one pause HMMs for Hungarian and 4 modalities and one pause HMMs for German. Statistics of segmentation and labelling are shown in Table 1. We used miscellaneous simple and complex sentences, and as it can be seen from Table 1, we placed about 1000 labels during the segmentation for Hungarian and 900 for German.

The energy (e_i) and fundamental frequency (f_{oi}) values were measured in a 25 ms time window, by 10 ms time frames. Measurement of fundamental frequencies were measured on the base of the Short-time Average Magnitude Difference Function (AMDF) We execute an octave filtering during processing of values of fundamental frequency, because the algorithm detecting the fundamental frequency could be wrong: it could jump an octave. Fundamental frequency curve is then linearly interpolated in logarithmic domain.

As the last step of pre-processing, the obtained energy- and fundamental frequency values were mean filtered in different time windows. These are the 5, 10, 20, 26, 30, 36, 40, and 50 frame numbers, multiplied by 10 ms frame time. The value averaged in a given sized interval will be the new value of the pattern located in the centre of the interval. We looked after the optimal time window length during our recognition test.

Besides e_i and f_{oi} values we computed 3-3 first and second derived ones on the basis of the size of three intervals both for the fundamental frequency and the intensity. This is the way to generate the parameter vector with 14 elements:

$$V_{jell} = \{f_{oi}, e_i, df_{oi}^{10}, d^2 f_{oi}^{10}, df_{oi}^{20}, d^2 f_{oi}^{20}, df_{oi}^{40}, d^2 f_{oi}^{40} de_i^{10}, d^2 e_i^{10}, de_i^{20}, d^2 e_i^{20}, d^2 e_i^{40}, d^2 e_i^{40}\}$$

The d, d^2 refer to the first and second derived, while the number being in the index after the derived ones refers to the size of the time interval used for the computation (in the unit of 10 ms).

2.2. **Training with different parameters**

We divided the processed database into two parts: we executed the training with the first part (529 sentences), and we made the testing with the second one (500 sentences). Sentences were mostly chosen randomly, but we paid attention that all labels to be recognized will be involved into both the train and test material. During the training we used the supra-segmental parameter vectors obtained from the speech files of the database with pre-processing, and the segmented and labelled data of the database for building the prosodic models of modality types.

2.3. **Testing**

Testing of the semantic level modality recognizer, it has two main processes: classification and. This process is shown in Figure 3. After the supra-segmental processing we used models describing the types of modality developed during the training, and some rules describing the connection of sentences, which we have been mentioned earlier. In the latter we specified grammatical rules that cover completely the cases emerging in the continuous speech regularly, with very few exceptions. Rules can give which

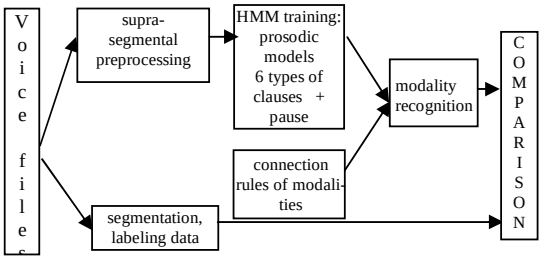

Figure 3. Flow diagram of testing

sentence or clause can follow a given sentence or clause, and which not; which sentences or clauses could be repeated, etc. Essentially it has an analogue role with the language model used in speech recognition, but as the number of possible connections is much less for sentences and clauses then for words of the vocabulary, we had given direct rules instead of statistical data.

We trained and used 6 clauses and a pause model to the recognition. We examined the correctness of recognition depending on the mean filter interval of supra-segmental parameter vectors (see in section 2.1.), and on the number of states of HMM clause/sentence models. We were searching for what kind of time resolution energy- and fundamental frequency-like characters need for an optimal classification of the different types of sentences.

The optimal number of clause and sentence HMM model states were also looked for. Evidently, there is a need for more than 3 states used in phoneme models, but we decided by testing, how many states were optimal to obtain the best classification results..

We examined the two factors together. The results are shown in Table 2. in case of Hungarian. Mean filter windows are given in the columns, the number of HMM states are written in the rows.

Table 2: Correct recognition of 7 different types of modality in (*Corr*) % in case of Hungarian

		Size of averaging interval by 10 ms frames							
		5	**10**	**20**	**26**	**30**	**36**	**40**	**50**
Number of HMM-stages	**5**	49,8	59,2	61,4	60,9	60,2	59,5	59,5	60,0
	11	66,2	68,7	69,0	70.1	69,7	68,4	70,60	66,3
	15	60,0	68,2	67,7	67,5	68,7	68,2	65,8	67,7
	19	-	-	67,9	65,3	66,3	64,6	64,6	61,2

Results can be seen in the cells of Table 2 indicating the percentage of correct modality recognition (*Corr*). We got the best results in case of 11 HMM states. Size of the mean filter window does not involve essentially the result between 100 and 400 ms. The best average recognition of modality was 70,6% (in *Corr)*. The confusion matrix for types of clauses can be seen in Table 3 for 11 state models and 40 frame mean filter
window. Rows of the matrix mean, what the original modality had been, and the columns mean, what the modality recognizer recognized. In case of certain types of sentences we had more samples for training and testing, and we got quite acceptable results: FF -50%, T – 83,3%, S – 74,8% and U –100,0%. The first three from them is worthwhile for recognizing types of modalities, and the last one is useful for detecting the sentence and clauses boundaries.

For German, where 4 modality and one pause model were only used for the recognition, we have got better results, the

Table 3. Confusion matrix for modality with state number 11 and mean filter interval of 40 frames. Overall *Corr* = 70,6%.

	S	**T**	**K**	**E**	**FF**	**N**	**U**	**Corr (%)**
S	**83**	11	7	4	2	3	1	74,8
T	4	**70**	0	1	2	0	7	83,3
K	3	3	**4**	0	0	0	2	33,3
E	1	1	0	**2**	1	0	1	33,3
FF	0	2	1	0	**5**	0	2	50,0
N	3	4	0	0	2	**4**	2	26,7
U	0	0	0	0	0	0	**125**	100,

correctness was 78,8 %, with 11 state models.

3. Evaluation

A HMM based semantic level modality recognizer have been developed, where the main aim was the optimization of the parameters of the HMM modality models. The recognizer with optimal adjustment of parameters gave better results, than it was expected, although the training set was relatively small, and we used an irregular database in the distribution of sentence types. We obtained the best recognition result, when the state numbers of HMM clause type-models were 11, and each state had 2 Gaussian components. With these adjustments the proportion of correctly recognized types of 6 modalities and the pauses for Hungarian is about 71%. S and T types can be detected with results of 75% and 83%, and the correct recognition of FF sentences can obtain 50% despite that we had only 52 sentences for training and testing. For German, the correctly recognized types of 4 modalities and the pauses is about 79%.

As it can be seen from our results, types of clauses trained with a large sized mass of data can be recognized with a good result, that is why it is suitable to process training material for the all types of sentences with similar quantity in the future. Development of a linguistic model for clauses with a large sized of mass of data on a statistical basis can also significantly improve the reliability of classification. The modality recognizer can be made more specific, if we build up clause models with training a database from spontaneous dialogues, because then the emotions (e.g. the prosodic features of exclamation and imperative sentences) could be more demonstrable, than on the basis of selections in the currently used read text databases.

We must continue working for improving the modality recognition of sentences and recognition of their boundaries. The purpose of our current article was to show, that the examination of these characteristics is efficient, and their involvement to the automatic speech recognition is useful.

4. References

[1] Ainsworth, William: Mechanisms of speech recognition. *Pergamon Press*, Oxford, 1976.

[2] Rabiner, L. R.: A tutorial on hidden Markov models and selected applications in speech recognition. Proceedings of the IEEE, 77(2):257-286, 1989.

[3] Young, S., Evermann, G., Gales, M., Hain, T., Kershaw, D., Moore, G., Odell, J., Ollason, D., Povey, D., – Valtchev, V., Woodland, P.: The HTK Book (for HTK Version 3.3.) *Cambridge University Engineering Department*, Cambridge, 2005.

[4] Becchetti, C., Ricotti, L. P.: Speech recognition, theory, and C++ implementation. *Fondazione Ugo Bordoni* – John Wiley, Rome. 1999

[5] Furui, S.: An overview of speaker recognition technology. In: Automatic Speech and Speaker Recognition (editors: Lee, C., Soong, F. K. Kuldip, K. P.), Kluwer Academic Publishers, 1996.

[6] Veilleux, N. M., Ostendorf, M.: Prosody/parse scoring and its application in ATIS. In: Human Language and Technology. Proc. of ARPA workshop, Plainsboro. pp. 335-340 , 1993.

[7] Kompe, R., Kiessling, A., Niemann, H., N¨oth, H., Schukat-Talamazzini E. G., Zottman, A., Batliner, A.: Prosodic scoring of word hypothesis graphs. In: Proc.of Interspeech 2005. Madrid,. pp. 1333-1336. 1995.

[8] Gallwitz, F., Niemann, H., N¨oth, E., Warnke, V.: Integrated recognition of words and prosodic phrase boundaries. In: Speech Communication, Volume 36, pp. 81-95, 2002.

[9] Vicsi K., Szaszák Gy.: Automatic Segmentation of continous Speech on Word Level Based Supra Segmental Features, Speech Technology, (pp. 363-370), 2006,

[10] Szaszák Gy., Vicsi K.. Using Prosody in Fixed Stress Languages for Improvement of Speech Recognition, Verbal and Nonverbal Communication Behaviours, COST Action 2102 International Workshop, Vietri sul Mare, Italy, March 2007, pp. 138-149

[11] Vicsi, K., Víg A.: First Hungarian speech database. Beszédkutatás '98. 163-177, 1998.

[12] Vicsi, K., Kocsor, A., Teleki, Cs., Tóth, L.: Speech database at a computer using environment. In Alexin Z., Csendes D. (eds.): II. Magyar Számítógépes Nyelvészeti Konferencia, Szeged. Szegedi Tudományegyetem Informatikai Tanszékcsoport, Szeged, 315-318., 2004.

[13] KIEL Corpus of read Speech, Volume I. Institut für Phonetik und digitale Sprachverarbeitung der Christian-Albrechts-Universitat zu Kiel, Dec. 1994.