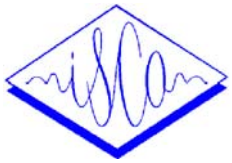

6th European Conference on
Speech Communication and Technology
(EUROSPEECH'99)
Budapest, Hungary, September 5-9, 1999

DEVELOPING THE DATABASE OF THE SPONTANEOUS SPEECH PROSODY CHARACTERISTICS

Jana Kleckova, Vaclav Matousek
Faculty of Applied Sciences, University of West Bohemia in Plzen
Univerzitni 22
CZ-306 14 Plzen
Czech Republic
kleckova@kiv.zcu.cz
Tel: +420 19 7491 177
Fax: +420 19 7491 213
http://www-kiv.zcu.cz/~kleckova

10.21437/Eurospeech.1999-177

ABSTRACT

The presentation deals with an experimental database of spontaneous-speech characteristics. The database is intended for a dialog system which have been developed in the Department of Computer Science at the University of West Bohemia within the framework of the Copernicus project. This system is assembled with a speech prosody module for processing the output of the acoustic phonetic module.. Proceeding that analysis the features obtained are evaluated using a neural network so that a type of the sentence can be determined. These results were verified in a number of experiments (1000 sentences). Having analysed the linguistic model we proposed further to extend the types of prosody characteristics and to create a database. Thus, besides the intonation analysis the prosody module also involves subroutines which can evaluate the pitch, both in a sentence and in a word, and the pause. The prosodic characteristics including the sentence are stored in the database and consequently exploited by the linguistic module as an additional information used for recognition and understanding the spontaneous speech. Processing the characteristics by usual methods of statistics the database can also be used to generate answers in the dialog system. The module was implemented in the C language an supported by the ORACLE database. For the user interface the environment SQL is suggested.

1. INTRODUCTION

Spoken word recognition involves the classification and the segmentation of a variable and continuous speech input. Different proposal have been made to characterize the relation between these two operations and the nature of the resulting units. A prototype of the dialog system developed in the Department of Computer Science carries continuously spoken human machine dialogs utilizing speech input and output techniques. The main components of the system prototype are: the speech input/output interface, the acoustic phonetic recognizer, the linguistic processor, and the dialog manager. In Czech language with its free-word-ordering intonation serves a critical information for the recognition and understanding system. For some sentences, the intonation is essential to determine the core of a communication, depending on a speaker who uses intonation to emphasive the meaning of a sentence. The design of the module for suprasegmental type processing is based on the partitioning the speech into sentences. In a such system prosodic attributes are determined by the acoustic--phonetic module. The time distribution of the voice energy and of the fundamental frequency is monitored within the period of a single sentence. The length of a pause as well as flags indicating word finality and lexical word accent are determined. Consequently, this information is used to associate the sentence with a certain type. The attributes determined by this procedure are used as the second input to the linguistic module.

2. PROSODY ATTRIBUTES

The ability of the listeners to identify correctly and almost instantly a word from amongst the tens of thousands of other words stored in their mental lexicon constitutes one of the most extraordinary human cognitive feats. The speech signal indeed presents a formidable challenge. Both the speech is variable (every word takes on a different phonetic shape each time it is produced – the existence of large numbers of a highly similar words in the lexicon makes this variability even more troublesome) and speech is continuous (unlike

written text, it contains no systematic *spaces* or reliable markers to indicate where word or utterance ends and the next one begins. The intonation often serves an information of a broad meaning nature. For example, the falling pitch we hear at the end of a statement in Czech such as „ *Vlak uz odjel. (The train has already gone*.)“ indicates that the utterance is complete. And on the contrary, the question „ *Vlak uz odjel? (Has the train already gone?*)“ in Czech equivalent has rising intonation. For this reason, falling intonation at the end of an utterance is called a terminal intonation contour. On the other hand, rising or unvaried level intonation, often indicates incompleteness. However, Czech sentences that contain question words like *kdy, co, kdo, jak (when, what, who)* usually do not have any rising intonation. It is like that the words in the question suffice to indicate that the answer is expected. The fact that rising or level intonations are correlated with incompleteness and falling intonation with completeness admits other utilizations of the intonation. One of them helps to make clear the interpretation of potentially ambiguous utterances. The prosody is a very complex subject. Besides the intonation the hierachy of pauses is very important. Pauses of standard length in the places of punctuation marks between syntactic units are felt as bizzare in the spontaneous speech. After several experiments have been treid out, a three-tier pause hierarchy seems acceptable in Czech.

Examples:

...(P3) Prosim Vas, (P1) muzete mi rict, (P2) kdy jede vlak do Prahy? (P3) ....
(Please, can you tell me, when the train to Prague is going. ....)

....(P3) Jak vidite,(P1) nezbyva nam mnoho casu.(P3}..
.(You see, there's not much time left.)

.....(P3)Uz se prilis nezdrzuj,(P1) Vaclave,(P2) a pokus se vlak dobehnout (P3)...
(Don't get stuck, Vincent, and try to run out the train.)

To make finer distinction of pauses would require to respect semantic relations of units in the dialog.

Pause	Duration of pause [ms] for speech rate	Classification of punctuaction marks
P1	8 - 10	{,}
P2	80 - 100	{- : }
P3	200 - 240	{; . ? !}

Table 1. Three-tier pause hierarchy

3. DATABASE OF PROSODY ATTRIBUTES

The design of the prosody module is based on the partitioning the speech into sentences. The sentences are processed using the following method:

- each sentence is divided into *n*-windows
- in each of the window is assigned with the voice energy (the first feature) and the fundamental frequency (the second feature)

The whole sentence is represented by *2n* features. The quality of the pattern recognition depends on the choice of the type and of the number of features. To classify sentences according to prosody, the prosodic characteristics must be computed and then considered as features. However, the features must be normalized and their number reduced to simplify recognition which then follows. Taking into account properties of the neural network employed in the recognition the number associated with a simple sentence must be the same for each one. The optimal number of features is proposed to be set 40 (in particular we consider 20 features energy and 20 features of frequency). As to our experience, a greater number than that proposed above does not improve the recognition. The attributes determined by this procedure form another input to the linguistic module. These results were verified in a number of experiments (1000 sentences). Having analysed the linguistic model we proposed further to extend the types of prosody characteristics and to create a database. Thus, besides the intonation analysis the prosody module also involves subroutines which can evaluate the pitch, both in a sentence and in a word, and the pause. The sentence pitch indicates an expression which may be crucial for identifying the core of the statement. Position of the sentence pitch is a matter of the statement realization, thus, being subjected to a context of the communication, or to a standpoint of the speaker. Therefore, it cannot be estimated, or even defined in advance. The prosodic characteristics including the sentence (features describing F0, energy , the length of the pause after and before the word, the speaking rate, flags indicating word finality and lexical word accent) are stored in the database and consequently exploited by

the linguistic module as an additional information used for recognition and understanding the spontaneous speech.

All the program equipment including the code SNNS

In the near future we shall use a neuronal network with different sets of prosodic features like duration of words, syllables and syllables nucleus, etc.

Type of the sentence	Corpus of sentences	Correct – Number	Correct- Percentage	Incorrect – Number	Incorect- Percentage
annoncement	100	71	71	29	29
question (query)	100	89	89	11	11
order	50	40	80	10	20
investigat.question	100	92	92	8	8
TOTAL	**350**	**292**	**83**	**58**	**17**

Table 2. Assignement of sentence with respect to the speaker's standpoint (read text)

Type of the sentence	Corpus of sentences	Correct – Number	Correct – Percentage	Incorrect - Number	Incorrect - Percentage
complex sentence - all sentences are significant	100	82	82	18	18
Complex sentence - 1. clause unsignificant	100	75	75	15	15
Single sentence	100	79	79	11	11
TOTAL	**300**	**256**	**78,6**	**44**	**21,4**

Table 3. Differentiating a complex sentence from a single sentence (read text)

[5] has been used in a number of experiments focused on the sentence evaluation. The list now follows:`

- assignment of sentences with the types with respect to the speaker's standpoint
- differentiating a question from other sentences
- differentiating an investigation question from other sentences
- differentiating a sentence construction from a single sentence.

4. EXPERIMENTAL RESULTS

The experiments reported in this paper were performed on a subset of Czech sentences - both read text and spontaneous speech. The sentences were generated using the ERBA templates [3]. To summarize the results is introduced in the tables Tab.2, Tab.3 , Tab.4, Tab.5 we can state that:

1. we are able to detect the types of the sentences.
2. the set of 80 features is has been probed (40 features energy and 40 features of frequency)
3. the experiment with read text brings the unconvincing .results.
4. a greater number than that proposed above does not improve the recognition.

5. CONCLUSION

It is currently that prosodic features have a very high significance for the dialog system. One important aspect of identifying elements of meaning (be it root morphemes, derivational or inflectional morphemes, words of various kinds or phrases) is their coding into segments, syllables and larger linguistic units like prosodic phrases. In the bottom-up process, acoustic cues in the speech signal are used by the listener in order to decode words. The aim of the experiments was twofold: First, the prosodic characteristics including the sentence (features describing F0, energy , the length of the pause after and before the word, the speaking rate, flags indicating word finality and lexical word accent) are stored in the database and second consequently exploited by the linguistic module as an additional information used for recognition and understanding the spontaneous speech. Processing the characteristics by usual methods of statistics the database can also be used to generate answers in the dialog system. The module was implemented in the C language an supported by the ORACLE database. For the user interface the environment SQL is suggested.

Type of the sentence	Corpus of sentences	Correct – Number	Correct- Percentage	Incorrect – Number	Incorect- Percentage
Annoncement	100	93	93	7	7
Question (query)	100	96	96	4	4
Order	50	40	80	10	20
Investigat.question	100	95	95	5	5
TOTAL	**350**	**324**		**26**	

Table 2. Assignement of sentence with respect to the speaker's standpoint (spontaneous speech)

Type of the sentence	Corpus of sentences	Correct – Number	Correct – Percentage	Incorrect - Number	Incorrect - Percentage
Complex sentence - All sentences are Significant	100	92	92	8	8
Complex sentence - 1. clause insignificant	100	90	90	10	10
Single sentence	100	99	99	1	1
TOTAL	**300**	**281**		**19**	

Table 3. Differentiating a complex sentence from a single sentence (spontaneous speech)

6. REFERENCES

[1] Kleckova, J., Krutisova,J., Matousek, V., Ocelikova, J.(1995). „An Automatic Creation of the Language Model for the Spontaneuos Czech Speech recognizer."*In: Proceedings of the European Conference EUROSPEECH'95, Madrid, September 1995*, 1185-1195.

[2] Kleckova, J. (1997) „Creation of the Language Model for the Spontaneous Czech Speech Recognizer: Sentence Formulas". *In: Proceedings of the 4th International Workshop on Systems, Signals and Image Processing, Poznan, May 1997,* 93-96.

[3] Kleckova, J. and Matousek, V.(1998) „ Detection of Sentence Types by the Integrated Prosody Module". *In: Proceedings TSD'98, Brno, September 1998,* 235 - 240.

[4] Palkova, Z.: Fonetika a fonologie cestiny. *Univerzita Karlova Praha, 1994.*

[5] Pal S.K., Dutta Majumder D.: Fuzzy sets and decision making approaches in vowel and speaker recognition. *In: IEEE Trans. Syst., Man, Cybern., vol.7, pp. 625-629,1977.*

[6]SNNS Stuttgart Neural Network Simulator (1995). User Manual, Version 4.1, cerven 1995.