Methods of Sentences Selection for Read-Speech Corpus Design

Vlasta Radová and Petr Vopálka

University of West Bohemia, Department of Cybernetics,
Univerzitní 22, 306 14 Plzeň, Czech Republic
radova@kky.zcu.cz

Abstract. In this paper methods are proposed which can be used to select a set of phonetically balanced sentences. The principle of the methods is presented and some experimental results are given. In the end of the paper the use of the proposed methods for the Czech read-speech corpus design is described in detail and the structure of the corpus is explained.

1 Introduction

One of the crucial problems that have to be solved when a speech recognition or a speech synthesis system is developed is the availability of a proper speech corpus for the system training and testing. The problem is usually solved in the following way: first a set of suitable sentences is selected from a database of phonetically transcribed sentences, next the set of the selected sentences is read by a group of speakers and, as the last step, the utterances are used to form a training or a test database [2], [3], [6].

The methods which are used to select sentences from the phonetically transcribed database can be divided into 2 groups. One of them consists of methods that enable to select sentences containing all phonetic events with approximately uniform frequency distributions. Such sentences are usually called *phonetically rich sentences* [4]. The other group includes methods that can be used to select "naturally" balanced sentences, i.e. sentences containing phonetic events according to their frequency of occurrence in natural speech. Such sentences are called *phonetically balanced sentences* [4].

Some ideas how to select a set of phonetically rich sentences were presented in our previous papers [7] and [8]. This paper deals with methods of phonetically balanced sentences selection. The principle of the methods together with some experimental results is presented in Section 2. Next, in Section 3, the use of the procedures in the course of the Czech speech corpus creation is described.

2 Methods of Phonetically Balanced Sentences Selection

The only way how to select the best set of sentences from a database of phonetically transcribed sentences is to form all possible sets and then to select the

best one. However, such a way may be too time consuming, especially when the number of sentences in the phonetically transcribed database and the number of sentences being selected are high. For that reason various procedures were proposed which are not so time consuming, however, they allow to select only a suboptimal set of sentences. The most known and used procedure is the *add-on procedure* [1], [4]. The procedure works with a phonetically transcribed text database and has the following 3 steps:

1. For each sentence of the phonetically transcribed text database a score S is computed that reflects how well the phonetic events contained in the sentence are represented in the up to now selected sentences.
2. The sentence with the best score is selected and moved to the list of the up to now selected sentences.
3. The steps 1 and 2 are repeated until the desired number of sentences is selected.

It is obvious that the most important problem in the procedure is the score S computation. To select a set of phonetically balanced sentences we propose the score

$$S = -\sum_{i=1}^{I} \left| \frac{m_i}{m} - \frac{n_i + n'_i}{n} \right| \tag{1}$$

where

$$m = \sum_{i=1}^{I} m_i \ , \tag{2}$$

$$n = \sum_{i=1}^{I} (n_i + n'_i) \ , \tag{3}$$

I is the number of distinct phonetic events that we wish to have in the selected sentences, m_i is the number of occurrences of the i-th phonetic event in the phonetically transcribed database, n_i is the number of occurrences of the i-th phonetic event in the up to now selected sentences and n'_i is the number of occurrences of the i-th phonetic event in the inspected sentence. Using this score, the sentence with the minimum score has to be selected in the step 2 of the add-on procedure.

The score (1) and the add-on procedure don't assure, however, that all phonemes will occur in the selected sentences. To overcame this disadvantage a *preselective procedure* has to be used before the add-on procedure. We propose the preselective procedure with the following 3 steps:

1. The sentence with the highest number of the distinct phonetic events which don't occur in the up to now selected sentences is selected from the phonetically transcribed text database and moved to the list of the up to now selected sentences. If two or more sentences can be selected in a moment the sentence which contributes mostly to the phonetical balance of the selected sentences is selected.

2. If some sentences can be excluded from the set of up to now selected sentences without decreasing the number of distinct phonetic events in the up to now selected sentences they are excluded and moved back to the phonetically transcribed text database.
3. The steps 1 and 2 are repeated until all phonetic events are present in the up to now selected sentences.

To test the work of the score (1) and the add-on procedure both without and with the preselection we tried to select a set of 40 phonetically balanced sentences from a set of 24 442 phonetically transcribed sentences. The sentences were selected with respect to the coverage of Czech phonemes [5], i.e. the Czech phonemes were regarded as the phonetic events in this experiment. Achieved results are presented in Table 1.

Table 1. The relative number of occurrences of particular phonemes [%]

Phoneme	in the primary set	in the set of 40 selected sentences		Phoneme	in the primary set	in the set of 40 selected sentences	
		without pre-selection	with pre-selection			without pre-selection	with pre-selection
e	9.2161	9.2322	9.2369	h	1.2451	1.0969	1.2048
o	7.9041	7.9525	7.9518	ee	1.1815	1.1883	1.1245
a	6.1693	6.2157	6.1847	sh	1.1033	1.0969	1.1245
i	6.1644	6.3071	6.1044	ch	1.0667	1.0055	0.9639
t	5.0174	5.0274	4.9799	x	0.9444	0.9141	0.9639
n	4.6839	4.5704	4.6586	uu	0.9185	0.9141	0.8835
ii	4.5716	4.5704	4.5783	f	0.8954	0.9141	0.8835
s	4.3688	4.3876	4.3373	tj	0.7744	0.8227	0.8032
v	3.9368	3.9305	3.9357	zh	0.7570	0.7313	0.7229
p	3.8056	3.7477	3.7751	rsh	0.7466	0.7313	0.7229
l	3.7974	3.7477	3.7751	ow	0.6589	0.6399	0.6426
r	3.7267	3.7477	3.7751	dj	0.4542	0.5484	0.4016
k	3.6420	3.5649	3.6948	rzh	0.4053	0.3656	0.4016
d	3.0798	3.1079	3.0522	g	0.3566	0.4570	0.4016
m	2.9849	3.1079	2.9719	ng	0.2104	0.1828	0.2410
nj	2.6900	2.7422	2.4900	aw	0.0386	0.0000	0.1606
j	2.6676	2.6508	2.7309	eu	0.0145	0.0000	0.0803
u	2.3687	2.3766	2.3293	dz	0.0124	0.0000	0.0803
aa	2.1479	2.1938	2.1687	oo	0.0113	0.0000	0.0803
z	2.1075	2.1024	2.0884	dzh	0.0019	0.0000	0.0803
c	1.5775	1.5539	1.6064	mg	0.0009	0.0000	0.0803
b	1.5741	1.5539	1.5261				

As the results show, using the add-on procedure without preselection the phonemes with a high relative frequency of occurrence in the primary set are

covered rather well in the selected sentences. However, several phonemes with a low relative frequency of occurrence in the primary set don't occur at all in the selected sentences. Using the add-on procedure with preselection all phonemes occur in the selected sentences, however, the phonemes with a low relative frequency of occurrence in the primary set are "overrepresented" in the selected sentences. This phenomenon can be however easily eliminated when more sentences will be selected.

3 Sentences Selection for the Czech Read-Speech Corpus

The goal of the Czech read-speech corpus is to provide enough continuous speech material for the development and evaluation of continuous speech recognition systems for Czech. We plan to record speech from at least 100 speakers from various regions of the Czech Republic. The texts to be read are selected from several Czech newspapers and have to satisfy several requirements. Each sentence must contain at least 3 and at most 15 words and have to contain no foreign words (i.e. words which are difficult to read for Czech people) and no numbers and abbreviations (since they may not be read identically by all speakers). Each speaker will be asked to read 150 sentences, where 40 sentences of the 150 are identical for each speaker. The remaining 110 sentences are carefully selected in order to satisfy several requirements. The sets of the 110 sentences from all speakers together will be used to train a Czech speaker-independent speech recognizer, the 40 sentences will be then used to adapt the recognizer to a particular speaker.

3.1 Selection of Adaptation Sentences

Our primary intention was to select the set of 40 adaptation sentences in such a way that the set will contain all triphones with approximately identical relative frequency. However, as the experimental results in [7] and [8] showed, this requirement was satisfied not very well for phonemes and the less it will be satisfied for triphones. For that reason we changed our primary intention and decided to select the adaptation sentences in such a way that they will contain triphones according to their relative frequency of occurrence in natural speech. To do it we used the add-on procedure described in Section 2 and the score (1). There were 2 additional conditions during the selection: no two adaptation sentences had to be identical and no sentence containing a triphone occurring only in that sentence had to be selected. The former condition is quite obvious, the latter one together with requirements posed on training sentences assures that no triphone which occurs in the phonetically transcribed sentences will be missing in the training sentences. Achieved results are given in Table 2. The sentences were selected from a primary set of 24 442 phonetically transcribed sentences containing 8 223 distinct triphones. The 40 selected sentences contain 1 492 distinct triphones what is only 18.14% of all distinct triphones occurring in the primary set. However, these 18.14% of triphones cover about 73.66% of

Table 2. Results of the adaptation sentences selection

Database	number of distinct triphones		covered text in the primary set
	absolute	relative with regard to primary set	
primary set	8 223	100%	100%
set of of adaptation sentences	1 492	18.14%	73.66%

the text in the primary database. Such a result can be regarded as a very good one since the 40 sentences will be used, as mentioned above, to adapt a speaker-independent speech recognizer to a particular speaker and therefore they should contain mainly the triphones occurring with a very high relative frequency in the natural speech.

3.2 Selection of Training Sentences

The set of training sentences has to satisfy two main requirements. It should contain all triphones occurring in the set of phonetically transcribed sentences and it should be phonetically balanced with respect to triphones. To do it we used the add-on procedure with preselection. The sentences were selected from the same set of phonetically transcribed sentences as the adaptation sentences, however, the adaptation sentences were already eliminated from the phonetically transcribed database. It means no adaptation sentence could be selected as a training sentence.

Using the preselective procedure 1 786 so called necessary sentences were selected which contain all triphones occurring in the set of 24 442 phonetically transcribed sentences. Remaining sentences were then selected to the necessary sentences using the add-on procedure. The number of the remaining sentences that had to be selected was given by the number of speakers in such a way that the total number of training sentences (including the necessary sentences) had to be 110 times higher than the number of speakers. In contradistinction to the adaptation sentences selection, a sentence could be selected several times during the training sentences selection. However, the number of repetition of each training sentence had to be at least 3 times lower than the number of speakers.

As mentioned above, a speaker-independent speech recognizer will be trained using the set of training sentences from all speakers together. However, each speaker will read only 110 training sentences. For that reason the whole set of training sentences was divided among particular speakers in such a way that each speaker will read 110 training sentences in total, no speaker will read a

sentence more than once and all speakers will read approximately equal number of sentences.

4 Conclusion

The paper deals with the problem of phonetically balanced sentences selection. Two iterative procedures have been presented which can be used to select a set of sentences that will contain phonetic events according to their occurrence in the natural speech. Both procedures have been tested on a primary set of 24 442 phonetically transcribed sentences from that a set of 40 phonetically balanced sentences was selected. In the end of the paper the use of the proposed methods for the Czech read-speech corpus design is described.

5 Acknowledgements

The work was supported by the Grant Agency of the Czech Republic, project no. 102/98/P085, and by the Ministry of Education of the Czech Republic, project no. VS 97159 and project KONTAKT no. ME 293.

References

1. Falaschi, A.: An Automated Procedure for Minimum Size Phonetically Balanced Phrases Selection. In: Proc. of the ESCA Workshop on Speech I/O Assessment and Speech Databases (1989) 5.10.1–5.10.4
2. Frasen, J., at al.: WSJCAM0 Corpus and Recording Description. Technical Report: CUED/F-INFENG/TR.192. Cambridge University, Engineering Department, Cambridge, UK (1994)
3. Gauvain, J.-L., Lamel, L.F., Eskénazi, M.: Design Consideration and Text Selection for BREF, a Large French Read-Speech Corpus. In: Proc. of the ICSLP (1990), 1097–2000
4. Gibbon, D., Moore, R., Winski, R. (eds.): Handbook of Standards and Resources for Spoken Language Systems. Mouton de Gruyter, Berlin New York (1997)
5. Nouza J., Psutka J., Uhlíř J.: Phonetic Alphabet for Speech Recognition of Czech. Radioengineering **4** (1997) 16–20
6. Paul, D.B., Baker, J.M.: The Design for the Wall Street Journal-based CSR Corpus. In: Proc. of the ICSLP (1992) 899–902
7. Radová, V.: Design of the Czech Speech Corpus for Speech Recognition Applications with a Large Vocabulary. In: Sojka, P., Matoušek, V., Pala, K., Kopeček, I. (eds.): Text, Speech, Dialogue. Proc. of the First Workshop on Text, Speech, Dialogue. Brno, Czech Republic (1998) 299–304
8. Radová, V., Vopálka, P., Ircing, P.: Methods of Phonetically Balanced Sentences Selection. In: Proc. of the 3rd Multiconference on Systemics, Cybernetics and Informatics to be held in Orlando, USA (1999)