A hungarian child database for speech processing applications.

Conference Paper · January 1999

Source: DBLP

CITATIONS
2

READS
33

3 authors, including:

Klara Vicsi
Budapest University of Technology and Economics
112 PUBLICATIONS **537** CITATIONS

SEE PROFILE

A HUNGARIAN CHILD DATABASE FOR SPEECH PROCESSING APPLICATIONS

Csatári, F. - Bakcsi, Zs. -Vicsi, K.
Technical University of Budapest, Hungary
csatari@ttt-202.ttt.bme.hu

ABSTRACT

This paper introduces a new Hungarian database containing spoken material recorded from children.
The aspects, which were taken into consideration under the selection of the speakers (and the composition of the speech database corpus), and the final content of the corpus is discussed. We described the method of the recording and the post-processing work on the recorded material. The paper also touches on the possible applications, in which the database is usable.
Different difficulties faced during the work, mainly arising from the age of our speakers, are reported.

INTRODUCTION

The processing of children's utterance is getting more and more important in the research and development of speech technology, mainly in speech therapy and in speech recognition [1],[2],[3]. Since a good speech database is needed in many speech processing application, collecting of a children's speech corpus is a very important task to do.
Our database reported in this paper is primary made for SPECO project of the INCO - Copernicus programme of European Commission, titled „A multimedia multilingual teaching and training system for handicapped children" [6], but we tried to make a widely usable speech corpus.

COMPOSITION OF THE TEXT MATERIAL

Our database is mainly collected to make the distance score evaluation of speech parameters between the Hungarian fricatives, affricates and vowels of right-speaking and speech handicapped children. The text material minimally had to contain all of these Hungarian phonemes in isolated form, in sound connections, in words and in sentences.
In sound connections, all vowels occur in concatenation with bilabial, alveolar and velar bursts, to present the coarticulation effects. In words, all examined speech sounds occurred in all sound positions and in all typical sound connections. One, two and three syllabic words were included. The sentences are designed to present the typical Hungarian intonation forms. You can see the structure of the text in Table 1.

In favour of the wide-ranging usability of the database (e.g. speech recognition), our aim was to provide as much as possible a phonetically rich material, including the most frequent Hungarian phonemes and sound connections. In an earlier detailed statistical examination [4], it was found, that half-syllable units give the most compact description of the phonological structure of Hungarian language and it is the reason why we tried to compose a half syllable rich material.
We analized the frequency of the occurred half syllables in the whole material. The result is to be seen in Table 2.
It is a very difficult task to construct a good children's speech database. Two aspects had to be considered. On the first hand the text material had to be large to represent a language as much as possible. On the second hand we had to be thoughtful of the age of our speakers. We can't use as long material as we want to, especially at the collection of children's speech. For example we had to take into consideration, that the spoken utterances mustn't be longer than 10-15 minutes, especially in case of 5 years old speakers. It is also a very important aspect, that the active vocabulary of 5 years old children is much smaller than the vocabulary of adults.

SPEAKER SELECTION

In our research we focused on 5-10 years old children.
The selection according the age is to be seen in Table 3.
As our speech database is primarily made for a teaching and training system for speech handicapped children, so it was important to study the voice of not only children with good, and average pronunciation, but also speech handicapped children. Therefore, we included in the database children with speech defects (approximately 40%), but those are only some examples.

Table 1. The structure of the text

Spoken utterance types		Examples
Sustained voices	vowels.	O, A:, E, e:, i, o, 2, u, y
	fricatives :, v	s, S, z, Z, f, v
Voice-connections	vowels with bilabial (p), alveolar (t) and velar (k) burst	pi: ti: ki, py: ty: ky, etc.
Digits	0-10	nullO (0), Ed' (1), kEtt2: (2), etc.
76 Words	monosyllabic dissyllabic trisyllabic	z2ld (green) t_sit_sO (cat) mu:zEum (museum)
29 Sentences with 5 different intonation forms	falling	O fA:n mo:kuS volt (There was a squirrel on the tree)
	quick falling	mEjik? (Which?)
	rising-falling	bOlA:Z hol vOn? (Where is Balázs?)
	quick falling-falling	kOti zEne:l? (Is Kate playing music?)
	floating	nEm, mA:r hOzOmEnt. (No, she already went home.)

Table 2. Occurrence of the 40 most frequent Hungarian a.) beginning b.) ending half syllables

a.)

Type of the half syllable	frequency in the 20th Hungarian prose	frequency in the text of the Hungarian child database
tO	2.76%	2.26%
tE	2.46%	2.63%
nE	2.41%	1.88%
mE	2.40%	3.38%
lE	2.28%	0.75%
nO	1.94%	1.13%
kO	1.91%	4.14%
O	1.76%	3.01%
mi	1.72%	0.75%
kE	1.68%	1.13%
lO	1.52%	1.50%
ho	1.41%	0.75%
hO	1.39%	0.75%
bE	1.32%	0.75%
rE	1.31%	0.75%
vO	1.31%	2.26%
mO	1.24%	1.50%
SE	1.21%	0.38%
ki	1.18%	1.13%
dE	1.18%	0.75%
zE	1.14%	0.75%
rO	1.06%	0.38%
te:	1.06%	0.38%
vE	1.02%	0.38%
ko	0.97%	1.50%
SO	0.96%	1.13%
ke:	0.96%	0.75%
ni	0.96%	1.13%
gO	0.94%	0.38%
lA:	0.93%	1.13%
sE	0.92%	0.38%
tA:	0.88%	0.75%
to	0.86%	0.75%
bO	0.85%	0.75%
mA:	0.82%	1.13%
jE	0.81%	0.38%
me:	0.81%	0.38%
E	0.79%	1.88%
ji	0.74%	0.38%
ti	0.74%	0.75%

b.)

Type of the half syllable	frequency in the 20th Hungarian prose	frequency in the text of the Hungarian child database
O	18.07%	16.17%
E	16.27%	6.02%
i	7.16%	7.14%
e:	5.17%	1.88%
o	4.95%	3.01%
A:	4.89%	2.26%
2:	1.67%	1.13%
2	1.51%	1.13%
o:	1.50%	2.63%
Em	1.49%	2.26%
u	1.33%	3.01%
El	1.27%	1.88%
Et	1.02%	0.75%
Ek	0.96%	1.50%
i:	0.93%	1.13%
e:S	0.92%	0.38%
En	0.87%	1.50%
ol	0.85%	2.26%
in	0.81%	0.38%
Er	0.71%	0.38%
iS	0.69%	0.75%
or	0.66%	0.75%
Ok	0.64%	0.38%
e:r	0.63%	0.38%
Ol	0.62%	0.38%
y	0.60%	1.13%
on	0.57%	0.38%
A:r	0.53%	1.50%
Ot	0.51%	0.38%
On	0.47%	1.13%
Eg	0.46%	0.38%
A:l	0.46%	0.75%
Es	0.44%	0.38%
u:	0.44%	1.50%
ok	0.43%	0%
Or	0.43%	0.75%
y:	0.40%	0.75%
ES	0.39%	0.38%
Os	0.38%	0.75%
A:S	0.35%	0.38%

Table 3. Age and gender distribution of speakers

age	children
5	10
6	10
7	18
8	16
9	6
10	12
total	**72**

All recorded children live in or near to Budapest. 5-6 years old speaker came from two local nursery schools, the 5-11 years old ones came from a local primary school. The hearing impaired children came from a special primary school for hearing impaired children in Budapest.

RECORDING

The recordings have been prepared in the anechoic chamber of the Acoustics Research Laboratory of the Department of Telecommunications and Telematics of the Technical University of Budapest.

The parameters of the anechoic chamber are the following: the size of the room is 125 cubic meters, the cutoff frequency is 90 Hz, the max. measuring distance is 5 meters and the noise level is at the hearing threshold.

The recording set-up contained a Monacor ECM-100 electret microphone and a Sony TCD-D7 DAT-recorder. The microphone was mounted on a microphone stand, and positioned 10 cm from the speaker's lips, 10 degrees off axis. However, in our case these values are only approximate, because it was often impossible to bring the children to keep their sitting position. The DAT was powered with batteries, to reduce power line influences.

The recordings were made with a sampling rate of 48 kHz and a resolution of 16 bit.

During the recording, both the operator and the speaker were sitting in the anechoic chamber. In this way we could achieve a more personal contact with children in the unfamiliar ambience. Also we had to take care, that the children do not change their position or that they do not touch the microphone or the stand.

Children, who could not read fluent yet, or syllabified the words, repeated the text by ear, the operator told the text first and the child repeated it. We asked the children to read the text, if it was possible. So the process became much faster. Sometimes only the sentences had to be spoken by ear, since even perfect-reading children had some problems with the intonation while reading.

The total recording time per speaker was approx. 10-15 minutes. We always recorded the whole conversation between the operator and the speaker, so the recordings contain not only the correct-spoken speech database corpus, but also badly pronounced or noisy utterances and some free conversation between the operator and the speaker. The correct utterances were selected during the post-processing work.

POST-PROCESSING

Further processing took place on a PC, so first the recorded material was transferred using a Turtle Beach Fiji PC sound card with digital I/O (this means with no quality-loss).

The post-processing works on the PC included the extraction of the voices, words and sentences to separate files, the selection of the correct, or if this was not possible, the best utterance, sample rate conversion to 22050 Hz, and backup at different stages of the processing works. The 22050 Hz sampled material is needed to ensure compatibility with the SPECO-programme.

Although the recordings were made on DAT-tapes, the database was archived on CD-ROMs.

On the CD-ROMs we recorded the materials as PCM wave files (Microsoft RIFF WAV).

Several archives were made during the making of the database:

- the original sound material (48 kHz, 10 CDs),
- the sample-rate-converted (22050 Hz) material with segment-markers and the extracted speech segments (4 CDs)
- the sample-rate-converted, extracted speech segments on 1 CD, to be used for the subjective monitoring (discussed later).

AREAS OF USE

Our database - as already mentioned - is mainly collected to give a good database for the work of the multimedia teaching and training system for speech handicapped children.

The database mainly designed to help finding examples for good, average and bad utterances, correlation between the subjective quality and computed quality (e.g. spectral distance) of the utterances and an etalon speaker, who could act as a reference e.g. in teaching applications. [5]

Moreover the database give good possibility for the acoustic and phonetic examination of child's speech.

Behind that, due to its phonetically richness, our database can be used to train respectively to test speech recognizers on children's voices.

CONCLUSIONS, FUTURE PLANS

The presented children database only a start-up of our collection work. Especially the extension of the number of the speakers is necessary, by normal speaking and speech handicapped children too. The increase of subset of speech defects would be very important for the SPECO project. During the next year we will expand this collection continuously.

ACKNOWLEDGEMENTS

The work has been supported by the Hungarian Scientific Research Foundation, and by the European Community as a part of the „SPECO"-Copernicus programme.

We would like to thank the following schools, making us possible to record their scholars:

„Török Béla" Hard of Hearing School in Budapest
Primary School on Érdi street in Budapest
Nursery School on Törcsvár street in Budapest
Nursery School on Zólyomi street in Budapest

REFERENCES

[1] A. Potamianos, Shrikanth, N. and Sungbok, L. (1997): „Automatic Speech Recognition for Children" EUROSPEECH '97, Vol. 5, pp. 2371-2374.

[2] S.M. Fosnot (1997) : „Vowel Development of /I/ and /U/ in 15-36 Month Old Children at Risk and not at Risk to Stutter" EUROSPEECH '97, Vol. 2, pp. 1051-1054.

[3] Sungbok, L. A. Potamianos, and Shrikanth Narayanan (1997) : „Analysis of Children's Speech : Duration, Pitch and Formants " EUROSPEECH '97, Vol. 1, pp. 473-476.

[4] Vicsi, - Vig, A. - Berényi, P.: „Text independent neural network/rule based hybrid, continuous speech recognition." COST 249 Contribution, February 1996., Kosice.

[5] Vicsi, K. - Csatári, F. - Bakcsi, Zs. „Distance Score evaluation of the visualized speech spectra at audio-visual articulation training" - EUROSPEECH.

[6] Vicsi, K. - Roach, P. - Öster, A. - Kaciè, Z. - Barczikay, P. - Sinka, I. : SPECO, a multimedia multilingual teaching and training system for speech handicapped children EUROSPEECH '99