PRELIMINARIES TO A ROMANIAN SPEECH DATABASE[1]

M. Boldea, A. Doroga, T. Dumitrescu, M. Pescaru

Department of Computer Science, "Politehnica" University
Blvd. Vasile Parvan 2, 1900 Timisoara, Romania
email: {boldea, ad48, td54, mt161}@cs.utt.ro

ABSTRACT

This paper presents the design and early recording stages of a Romanian speech database to be used for development of both speech recognition and speech synthesis systems.

The recognition part is built around a core patterned after the EUROM_1 [4] design, so that an as good as possible compatibility to exist with this, and includes both read and semispontaneous speech. The synthesis part consists of a read speech corpus from which diphones are to be extracted to build concatenation-based TTS systems, and read material to serve as benchmark data for the administration of a Romanian version of the Modified Rhyme Test [2].

1. INTRODUCTION

As a series of technologies are demonstrated to be adequate for real life applications of voice interfaces, speech databases become a key component in expanding these technologies to new languages.

Our work aims at gathering a database of Romanian speech able to provide for: acoustic phonetics studies; text-to-speech systems building and evaluation; automatic speech recognizers development and test. Although more directly concerned with the last two objectives, we hope that the availability of this speech corpus would also stimulate basic research in Romanian acoustic phonetics and phonology, the scarcity of which had been unpleasantly felt by us during the design of this database.

The main stages in the database development completed or entered so far are: 1) corpora design, to specify the type and contents of the material, detailed in Section 2; 2) speech signal recording, presented in Section 3; 3) post-recording processing (Section 4). Future work is outlined in Section 5.

2. DATABASE DESIGN

In order to develop voice-based computer interfaces, a first step is the collection of as much speech signal as possible, first of all to be able to model its variability along such dimensions as speakers' sex, age, cultural background, accent etc.

[1]This work has been supported by the European Commission through contract COPERNICUS-1304/1994 and the Romanian Ministry of Education through contract 4004/41B/1995.

Starting from this, the goals set forth are to provide speech material to be used for both acoustic modeling in speaker independent automatic speech recognition systems development and test, and text-to-speech systems building and evaluation. During this stage there appeared to be necessary also more detailed knowledge in Romanian acoustic phonetics and phonology, which hopefully will also result from studies to be carried out using this database.

2.1. The recognition part

This part is intended to be used for training, development and evaluation of continuous speech recognizers at the phoneme and word levels using speaker independent phoneme models, with some context dependence modeling capability. Although initially planned to contain read speech based on prompts extracted from newspaper or similar texts, which would have allowed for its use for up to speaker independent very large vocabulary continuous speech recognition, the difficulties with obtaining text archives of adequate size led us to consider other alternatives.

The final solution chosen was to build it around a core patterned after the EUROM_1 database [4] already used for other 11 European languages, adopted as such for another group of five Central and Eastern European languages for which speech databases are developed now within a European Commission funded COPERNICUS project, and which contains for each language:

- 40 short passages of five thematically connected sentences with themes common to all languages;
- sentences composed to compensate for the phoneme frequency imbalance in the passages, where necessary;
- 100 selected integer numbers between 0 and 9999 such that the main phonotactic possibilities of these numbers be covered;
- C(C)VC(V) material in isolation and in context;
- five pairs of context words for use with the C(C)VC(V) material.

The core itself contains 40 short **passages** translated and adapted from the English version of the EUROM_1 passages: the translations were made as phonemic transcriptions using a

Phoneme	Clusters	60 sp.	80 sp.	100 sp.
h	70	457	605	751
dZ	72	474	630	785
Z	72	492	650	804
o_X	76	556	732	904
w	81	598	792	970
g	98	742	968	1183
b	108	850	1108	1349
z	112	813	1068	1316
e_X	119	969	1262	1542
f	129	1015	1326	1601
ts	132	1025	1348	1642
v	147	1097	1442	1782
i_0	177	1252	1653	2039
tS	181	1324	1739	2145
S	195	1424	1853	2273
1	222	1798	2349	2852
j	293	2227	2935	3613
m	362	2887	3780	4643
p	363	2759	3591	4396
o	377	2912	3835	4727
d	384	2987	3897	4777
k	437	3361	4392	5395
l	456	3613	4728	5790
s	478	3757	4919	6017
@	508	3870	5073	6198
u	598	4777	6231	7601
i	677	5286	6912	8437
n	707	5513	7224	8812
t	711	5443	7141	8750
r	803	6054	7950	9756
a	1226	9213	12079	14823
e	1233	9330	12232	15034

Table 1: Total number of expected occurrences for every Romanian phoneme in: the ten extended clusters; extended clusters and personal sentences to be read by 60, 80, and 100 speakers.

	Entropy	Distinct words	Total words
Cluster 0	4.51	177	272
Cluster 1	4.54	170	244
Cluster 2	4.49	182	256
Cluster 3	4.49	178	267
Cluster 4	4.48	161	219
Cluster 5	4.51	174	238
Cluster 6	4.51	171	238
Cluster 7	4.55	177	246
Cluster 8	4.53	166	239
Cluster 9	4.55	188	259
All clusters	4.53	1160	2478

Table 2: Phonemes distributions entropies and numbers of distinct and total words (per extended cluster and globally)

notation based on SAMPA [7], such that a statistical analysis could be carried out at the phonemic level and a heuristically-based automatic clustering procedure used to group them in ten clusters of four passages each, with as even as possible phonemic distributions.

Further, a few **filler sentences** built as phonemic transcriptions in a computer aided interactive manner were added to each cluster to make sure that all phonemes have a minimum number of seven occurrences in each cluster, minimum chosen such that an automatic segmentation and labeling procedure could be run on these extended clusters with enough occurrences of every phoneme in order to obtain good estimates of the phoneme boundaries [5]. The global phoneme statistics for the extended clusters are given in Table 1. The informational divergences of phonemic distributions evaluated for all extended clusters had an average of 0.032 and a standard deviation of 0.0075, with values between 0.02 and 0.05, which we consider to be a good proof of their similarity. This is also demonstrated by their entropies and distinct and total number of words (Table 2).

The **numbers** part of EUROM_1 was redesigned to reduce its size while preserving its coverage of the phonotactic possibilities of the Romanian integers between 0 and 9999. As a result, it consists now of 26 numbers, most of them composed of three and four digits. It is intended as training and test material for recognizers on a very small continuous speech recognition task, but with many interesting variations in pronunciations, and for text dependent speaker identification.

The last part to ensure the EUROM_1 compatibility comprises **CVC type words** to be read in isolation and in five controlled contexts by ten respectively two of the first 60 speakers corresponding to the Few Talker and Very Few Talker sets in EUROM_1 [4]. The envisaged use of this part is in coarticulation studies and speech recognizer assessment [6].

Extensions added to this core are:

- a set of four **initialization sentences**, built as phonemic transcriptions in a computer aided interactive manner, such that each of them contain at least one occurrence of every Romanian phoneme, to be recorded by all speakers, manually labeled and used to initialize phoneme HMMs for an automatic segmentation and labeling of the database; they can also provide material for text dependent speaker identification;
- a set of **personal sentences** automatically selected using a greedy algorithm from a text corpus in order to provide greater variability and ensure an as large number of diphones occurrences as practically feasible given the available data; a first step before the selection itself was the development of a pronunciation dictionary (about 24000 words) for the whole text corpus, and the manual extraction of sentences suitable to be read; each speaker will read a

different subset (between three and seven) of these sentences; the last three columns in Table 1 include the expected number of phonemes to result from these sentences;

- the **semispontaneous speech** part is intended to facilitate studies on differences from read speech and developments of some simple interactive voice response systems and contains recordings of a few **personal information items** for each speaker (**name - spoken and spelled**; series and number of the personal **identification document** - two letters + six digits; **telephone number** - six to nine digits; **birth date**; **address**) and of the Romanian **alphabet**.

The ten extended clusters will be used to collect data from groups of ten speakers of each sex, such that the data collection will proceed in increments of 20 speakers, with a minimum of 60 speakers planned for compatibility with EUROM_1 [4], an intermediate 80 speakers checkpoint, and an intended 100 final speakers. Every 20 speakers increment is foreseen to consist of four speakers in each of five age groups (under 20, 20-29, 30-39, 40-49, over 50 years), although this might be subject to reevaluation during the data collection. No effort is being made to obtain any accents coverage.

2.2. The synthesis part

The purpose of this part is twofold:

- to provide signal for speech synthesis in TTS systems; given its simplicity, the concatenation method has been envisaged, with diphones as the first units to be experimented with; consequently, a series of items has been recorded from which the diphones are to be excised;
- to make possible the administration of synthesized speech intelligibility tests at the segmental level, for which a Romanian version of the Modified Rhyme Test [2] has been devised, and the words in its lists have been recorded to be used as benchmark data [3].

Given its small dimension, compared to the rest of the database, we expect this part to be re-recorded as more experience is gained with its use.

3. RECORDING

The data collection is done using a SESAM workstation equipped with an OROS-AU21 board and running the EUROPEC corpus recording software [8] in standard SAM format files.

The recordings take place in a soundproof room using a SONY ECM-44B electret condenser microphone placed about 25 cm from speaker's mouth and connected through a preamplifier to the SESAM workstation situated in an adjacent room.

Although the configuration allows for prompts presentation using a computer display, this solution was abandoned due to acoustic noise being produced by the deflection coils, and prompt texts are read by speakers from paper listings. Instructions are given through an intercommunication system. Except for the initialization sentences, where pronunciations as close to the standard form as possible are sought, no restrictions are imposed on speakers' pronunciations except for word deletions, insertions or substitutions.

Speakers' personal data are collected according to the EUROPEC speaker description format [8].

4. POST-RECORDING PROCESSING

This includes:

- a quality check for every speech signal file as resulted from the recording session;
- the high pass filtering of every speech signal file using an 1661 coefficients FIR filter with a cutoff frequency of 72 Hz in order to remove mains related noise;
- a new quality check for every filtered speech signal file;
- the manual labeling of the four initialization sentences.

The quality checks, based on the NIST SPQA package, verify: the DC bias; the existence of signal clipping; the signal and noise levels; the signal-to-noise ratio; the presence of mains related noise components.

For the first 24 speakers (16 male, 8 female) from the recognition part, the average SNR after filtering is 50.38 dB, with a standard deviation of 2.69 dB, and extreme values of 43.5 and 60.25 dB.

5. FUTURE WORK

Besides continuing the data collection, the manual segmentation and labeling of the diphones database is under way, as well as that of the initialization sentences. Work is in progress towards the automatic segmentation and labeling of as much data as possible using an HMM-based system [1], to be followed by a manual label verification and correction stage.

The manual labeling reveals potential problems with the presence of the devoiced /i_0/ vowel, whose occurrences are very difficult to locate (examine e.g. Fig. 1 and [SOUND A972S01.WAV], from which the signal was extracted), which might make necessary the recording of a special corpus to study its properties.

6. ACKNOWLEDGMENTS

Thanks to those members of the previous SAM projects who made available the EUROPEC corpus collection software and the EUROM_1 documentation, to Mark Huckvale of University College, London for SFS, and to all COPERNICUS-1304 project

partners who advised us on database design and collection issues. Thanks are also due to our colleague Nicolae Robu who helped with the soundproof room, as well as to all speakers who freely donated their speech so far.

7. REFERENCES

1. Brugnara, F., Falavigna, D., and Omologo, M., “Automatic segmentation and labeling of speech based on Hidden Markov Models”, Speech Communication, Vol. 12, no. 4 (August 1993), pp. 357-370

2. House, A.S., Williams, C.E., Hecker, M.H.L., and Kryter, K.D., “Articulation-Testing Methods: Consonantal Differentiation with a Closed Response Set”, J. Acoustic. Soc. Amer., Vol. 37, 1965, pp. 158-166

3. Logan, J.S., Greene, B.D., and Pisoni, D.B., “Segmental intelligibility of synthetic speech produced by rule”, J. Acoustic. Soc. Amer., Vol. 86(2), 1989, pp. 566-581

4. The SAM Projects, “EUROM - A spoken language resource for the EU”, Proc. EUROSPEECH’95, pp. 867-870

5. Schmidt, M.S., and Watson, G.S., ”The evaluation and optimization of automatic speech segmentation”, Proc. EUROSPEECH’91, pp. 701-704

6. Steeneken, H.J.M., and van Velden, J.G., “Recognizer assessment by means of CVC-words as available in the EUROM-1 data-base”, SAM-TNO-040, TNO-Institute for Perception, Soesterberg, The Nederlands, October 1991

7. Wells, J.C., “Computer-coding the IPA: a proposed extension of SAMPA”, University College, London, 1995

8. Zeiliger, J., and Serignat, J.F., “EUROPEC Software V4.1 User’s Guide”, Institute de la Communication Parlee, Grenoble, France, 1991

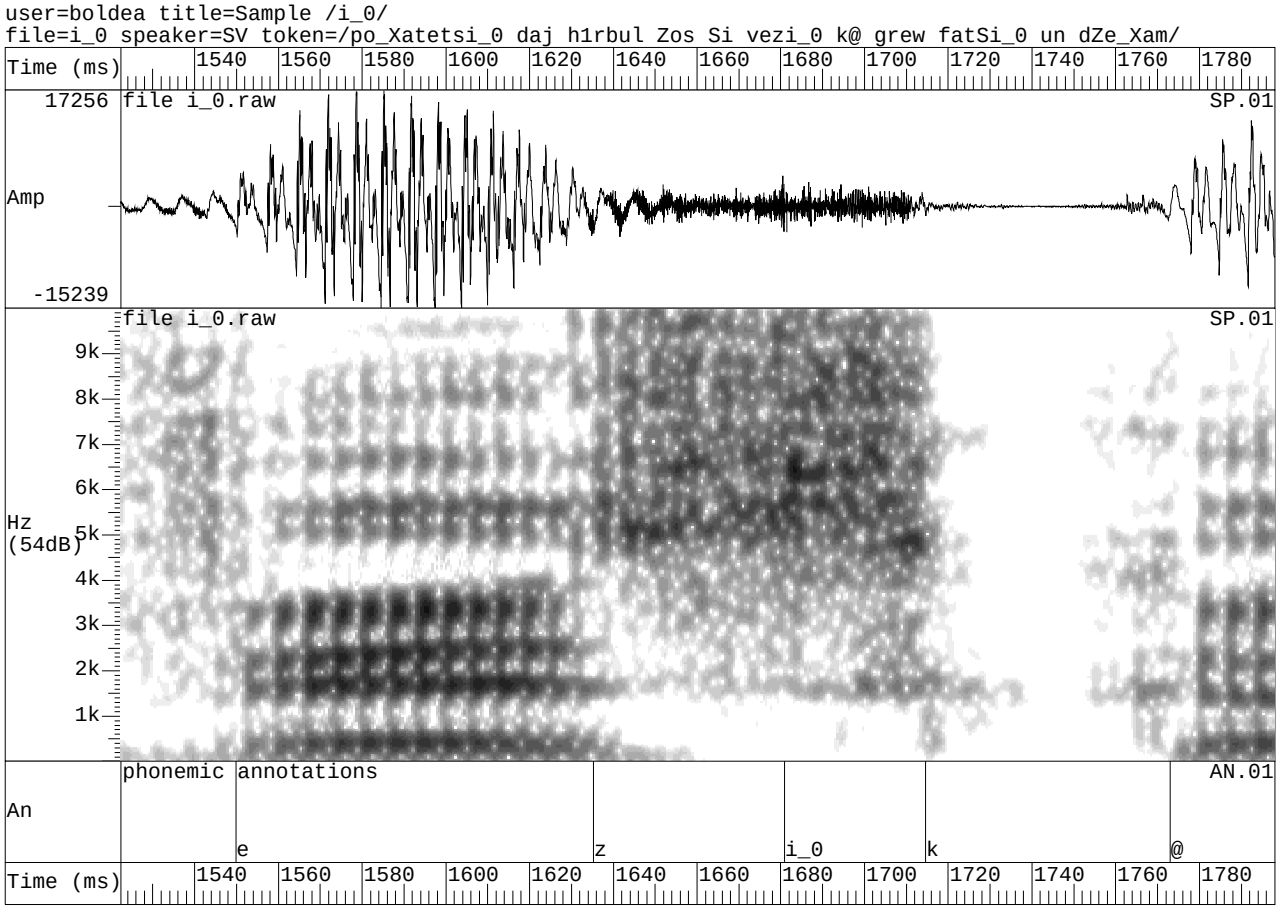

Figure 1: Sample occurrence of a devoiced /i_0/ vowel. Listen [SOUND A972S01.WAV] for the whole sentence.