*Proceedings of the First International Conference on Language Resources and Evaluation - LREC. Workshop on Speech Database Developement for Central and Eastern European Languages, Granada, Spain, 1998*

# Design, Collection, and Annotation of a Romanian Speech Database

**Marian Boldea, Cosmin Munteanu, Alin Doroga**

Department of Computer Science, "Politehnica" University of Timisoara
Blvd Vasile Pârvan 2, 1900 Timisoara, România
email: {boldea,cosmin,alin}@ear.utt.ro

**Abstract**

Speech databases are essential resources for the acquisition of linguistic knowledge and speech technology developments, and both can be facilitated if the collected signal is accompanied by some form of annotation. This paper gives an overall view on the design, collection, and annotation of a Romanian speech database including over 10 hours of speech from 100 speakers, labeled in part at the broad phonetic level.

## 1. Introduction

Started as part of the COPERNICUS BABEL project, the construction of a Romanian speech database (Boldea et al., 1996) including both read and semispontaneous speech from 100 speakers is now in its final stage, and this paper will give an overview of the whole process. In section 2, details concerning database design (material to be recorded, speaker population) are presented, followed in section 3 by a description of the data collection proper. Signal labeling at the broad phonetic level is treated in section 4, and section 5 describes present status and work yet to be done before the database will be ready for distribution. Finally, sources of support are acknowledged.

## 2. Database Design

Spoken language corpora as research tools should be designed and collected defining and keeping in mind as clearly as possible what their use would be (EAGLES, 1994), and we chose as our primary objective the development of speaker independent continuous speech recognition in Romanian, which implies a large speaker population, and a controlled number of occurrences for each of a predefined set of acoustic modeling units, based on having speakers read specially designed prompts, so that model parameters could be estimated reliably. In our case, as this is the first Romanian database of its kind, and there was no previous experience with modeling units for continuous speech recognition in Romanian, we decided that the database allow for acoustic modeling at the phoneme level.

Since pre-normalization and standards were among the objectives of the COPERNICUS programme, in order to comply with them we started from the existing EUROM-1 database (Chan et al., 1995) with special emphasis on: an integrated redesign of various components as found in EUROM-1 (read passages, filler sentences, numbers, CVC words isolated and in contexts) to obtain a more systematic satisfaction of their aims; adding new read and semispontaneous materials; a speaker population of minimum 60 persons with a uniform age and sex group distribution, extensible beyond this limit, and structured in Many Talkers, Few Talkers, and Very Few Talkers sets similar to those in EUROM-1.

### 2.1. Recording prompts

Many of the recording prompts were built around adapted translations of the **40 passages** in the English version of EUROM-1, grouped in **10 clusters** of 4 passages each through a heuristic procedure seeking a uniform a priori number of occurrences across clusters for every phoneme (fig. 1); in this procedure, the deviation from the uniform distribution allowed for a phoneme in one cluster was adaptively changed until a solution was found, and in our case:

$$\Delta = \frac{73(total\_no[phoneme] - 15)}{\max(total\_no) - \min(total\_no)}$$

where *total_no* is an array holding the total number of occurrences for every phoneme.

```
list phonemes in ascending order of frequency
foreach phoneme
  list passages in ascending order of phoneme occurrences
  list clusters in descending order of phoneme occurrences
  while (not all clusters OK) // cluster OK with phoneme if
     // contains total_no[phoneme]/NO_CLUSTERS ± Δ
     // occurrences of phoneme
     foreach cluster
       if (cluster not OK)
          allocate first_passage to cluster // to maximize the
          // number of occurrences of phoneme in cluster ;
          // fails if passage list empty or cluster full
          delete first_passage from the list
       endif
     endfor
  endwhile
endfor
```

Figure 1: Heuristic passages clustering procedure

Because in the basic clusters some phonemes were poorly represented, each of them was extended with 2 or 3 **filler sentences**, manually built using a special

editor indicating needed phonemes, to raise to a minimum level the expected number of occurrences of the least frequent phonemes, and ten **extended clusters** were obtained.

The minimum number of occurrences per extended cluster for a phoneme was set to seven to facilitate signal labeling using automatic procedures (Schmidt & Watson, 1991), and for the same purpose four **phonemically compact sentences**, common to all speakers, were added, with the resulting signals to be labeled by hand and used to initialize segment models.

To increase phonetic variation and provide for some context dependent modeling at the diphone level, about 550 **individual sentences** (between 3 and 7 distinct sentences per speaker) were added from a text corpus by a greedy automatic selection procedure.

Other read materials similar to those in EUROM-1 were included for performance and diagnostic evaluation of speech recognizers: integer **numbers** between 0 and 9999, which in EUROM-1 are the same 100 for all languages, were replaced by a reduced set of 26, checked to satisfy the phonotactics coverage intended by EUROM-1, and the **CVC words**, in isolation and in controlled **contexts**, were adapted to the Romanian phonological system.

Finally, **semispontaneous materials** were planned to be collected by requests for very simple personal data (**names** - spoken and spelled; **ID code** - two letters and six digits; **telephone number**; **birth date**; **address**) to study speaking style differences and develop specific applications, and a reading of the **Romanian alphabet** was included for comparisons with names spelling and ID code letters pronunciations.

### 2.2. Speaker population and prompts distribution

For compatibility with EUROM-1, a minimum number of 60 speakers, from which 10 Few Talkers and 2 Very Few Talkers, was planned to be recorded, with an even distribution across sex and age groups (under 20, 20-29, 30-39, 40-49, 50 and over). This lead to splitting the prompt passages in 10 clusters, which in turn allowed each extended cluster to be planned for reading by at least three speakers of each sex, and additional speakers be recorded in 20 speakers increments.

Every speaker was planned for a recording session of semispontaneous materials, the Romanian alphabet, one extended cluster, 3 to 7 individual sentences, four phonemically compact sentences, and the integer numbers.

From the first 60 speakers, ten were selected in the Few Talkers set, one per sex and age group, and planned to record additional specific materials (isolated CVC words, four new extended clusters and four repetitions of the numbers) so that materials similar to those in EUROM-1 be collected.

The Very Few Talkers set includes one male and one female speaker from the Few Talkers set, whose specific additional material consist of CVC words in contexts and the context words.

## 3. Data Collection

Recordings took place in a sound treated room using the EUROPEC data collection software (Zeiliger & Serignat, 1991) running on a PC-compatible computer placed in an adjacent room, and equipped with an OROS AU-21 A/D-D/A conversion board.

Through fonts redefinition and appropriate encoding, provisions have been made for Romanian diacritics both in EUROPEC messages and prompt texts, but eventually, although EUROPEC allows for recording instructions and prompts presentation on computer monitors, paper listings were used for prompts, and an operator-controlled intercom to instruct speakers, in order to avoid acoustic noise produced by deflection coils.

A SONY ECM-44B electret condenser microphone, placed about 25 cm from speaker's mouth, 30 degrees off axis, was connected through a fixed-gain preamplifier to the OROS AU-21 board, whose sampling rate was set to 20000 Hz with 16 bit per sample.

Every resulting signal file, corresponding to a prompt, is in raw PC (little-endian) format, and accompanied by configuration and description (item type for semispontaneous materials and alphabet, orthographic transcription for the rest) files in SAM (EAGLES, 1994) formats.

For each speaker, the (first) recording session started with an agreement being signed and personal data collected and introduced in a global speakers description file, followed by instructions and training using semispontaneous materials and alphabet reading, with the resulting files discarded, so that actual recordings were done trying to minimize pronunciation alterations due to speaker stress or speaking style changes, in the sequence: semispontaneous materials, read alphabet, passages, filler sentences, individual sentences, phonemically compact sentences, numbers.

The Few Talkers continued with isolated CVC words, and the Very Few Talkers - with CVC words in isolation and in contexts. Each of the ten Few Talkers recorded four additional sessions at least two weeks apart, in which one new extended cluster and the same numbers were read.

With the exception of the phonemically compact sentences, for which pronunciations as close to the standard as possible were required, and corrections of word deletions, insertions, or substitutions, no restrictions were imposed on speakers.

In order to preserve consistent quality along all the recording period, the signal files collected in a recording session were checked immediately for a number of quality parameters: DC bias, signal clipping, signal and noise levels, signal-to-noise ratio, and mains-related noise components.

Recordings stopped at 100 speakers, and three CD-ROMs were written holding all the collected data.

## 4. Signal annotation

Besides the annotation already available for each signal file in its associated description file generated during the recording session, the phonetic labeling of some of the signals was planned initially in the BABEL project, and additional provisions have been included early in the database design phase in order to facilitate its extension to all signals using automatic segmentation and alignment techniques.

### 4.1. Speech signal labeling

Whatever the use, a segmental labeling of the collected signals increases their value, and from the acoustic, narrow, and broad phonetic labeling levels, the last was chosen due to its being the most economical, the most appropriate for speech recognition systems training and evaluation (Barry & Fourcin, 1992), and because it offers the highest labeling reliability (Eisen, 1996), in that transcriptions consistency across labelers, for the same speech signals, is maximized, although boundary placement consistency is about the same with that at the narrow phonetic level.

Done by hand, speech signal labeling is extremely time consuming, and various approaches were tried to make it (at least in part) automatic.

The first large acoustic phonetic, and probably most known, reference speech database, TIMIT (Garofolo et al., 1993) was labeled using a semiautomatic procedure (Zue & Seneff, 1988) consisting of a manual quasi-phonemic transcription stage, an automatic speech signal segmentation and label alignment using acoustic-phonetic rules, and a final correction by hand of label identities and segment boundaries based on listening and visual examination of speech signal waveforms and spectrograms.

More recently, as Hidden Markov Models (HMM) were established as fundamental tools in speech technology, HMM-based automatic segmentation and alignment became dominant (Ljolie & Riley, 1991; Brugnara, Falavigna & Omologo, 1993; Kipp, Wesenick & Schiel, 1996), usually using label networks generated automatically from orthographic transcriptions by TTS components, and including pronunciation variants specified by phonological rules.

Because this is the first Romanian speech database of its type, possibly to be established as a reference, and no phonological rules were available for network generation, we chosen a labeling methodology similar to that used for TIMIT, i.e. manual transcription, HMM-based automatic segmentation and label alignment (fig. 2), and manual verification and correction.

### 4.2 Signal transcription

Although final label files will be generated in SAM format (EAGLES, 1994) using SAMPA symbols (Wells, 1995), the symbol set used for transcription is composed exclusively of single lower and upper case ASCII characters (table 1).

Signal transcription was based on listening and visual examination of waveforms, and included continuous speech phenomena (assimilations, elisions, epentheses, etc.)

### 4.3. Automatic alignment

As an outcome of the transcription process, each signal file is accompanied by a transcription file, and an automatic segmentation and label alignment can be done by iteratively training segment (phoneme and silence) HMMs and using them for a Viterbi segmentation of the signals (Gauvin & Lamel, 1992; Rabiner & Juang, 1993).

| ASCII | SAMPA | Example word(s) |
|---|---|---|
| i | i | si (and) |
| I | C | azi (today) |
| e | e | deget (finger) |
| y | 1 | în (in), când (when) |
| @ | @ | dacă (if) |
| a | a | lac (lake) |
| u | u | nu (no) |
| o | o | cot (elbow) |
| j | j | ieri (yesterday) |
| E | e_X | deal (hill) |
| w | w | nou (new) |
| O | o_X | coate (elbows) |
| p | p | cap (head) |
| b | b | bere (beer) |
| t | t | timp (time) |
| d | d | dop (cork) |
| k | k | camera (room) |
| g | g | gluma (joke) |
| T | ts | țara (country) |
| C | tS | cer (sky), ceai (tee) |
| G | dZ | gem (jam) |
| f | f | fata (girl) |
| v | v | vin (wine) |
| s | s | sare (salt) |
| z | z | zbor (flight) |
| S | S | șapte (seven) |
| J | Z | joc (game) |
| h | h | harta (map) |
| m | m | mic (small) |
| n | n | nas (nose) |
| l | l | lapte (milk) |
| r | r | rosu (red) |
| _ (underscore) | ... | silence |

Table 1: ASCII and SAMPA label symbols

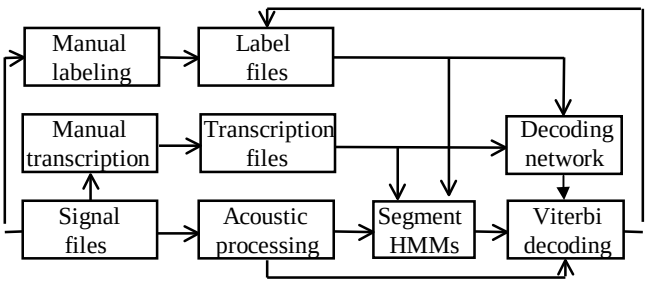


Figure 2: Semi-automatic segmentation and alignment

A first step towards this was the manual labeling of the four phonemically compact sentences, common to all 100 speakers, to be used for segment HMMs initialization.

Three-state left-to-right sex-dependent HMMs with Gaussian density mixture output probability functions were used. To obtain a more rapid convergence of the training process, the phonemically compact sentences labeled manually were used for a segmental K-means initialization of mixture parameters followed by a Baum-Welch reestimation of both mixtures and transition probabilities, and a concatenated training used all available signal and associated transcription files.

To obtain a good time resolution, the acoustic processing used frames 12.8 ms long, spaced at 5 ms, from which 26-dimensional vectors including 12 autocorrelation LPC liftered cepstral coefficients, log energy, and their first derivatives, were computed.

Automatically aligned label files were obtained through a Viterbi decoding of each signal file guided by a network generated from the associated transcription.

### 4.4. Label verification and correction

Once automatically generated label files were available, they were verified and corrected based on listening and visual examination of signal waveform, spectrogram, and labels. To ensure consistency, a display at a sex-dependent constant resolution was used, and rules in cases of arguable boundary placements.

## 5. Present Status and Future Work

More than 10 hours of speech were recorded from 100 speakers with a uniform sex and age group distribution, and three CD-ROMs were produced.

For all 100 speakers, the phonemically compact sentences, one extended cluster, and the individual sentences, amounting to a total of about 3200 sentences, were already labeled at the broad phonetic level.

Future work is intended to complete the database labeling and produce CD-ROMs including label files and documentation.

## 6. Acknowledgments

This work has been supported by the European Commission through contract COPERNICUS 1304/1994, the Romanian National University Research Council through grant 354/1996, the Romanian Academy through grant 136/1997, and the Romanian Ministry of Research and Technology through contract 3019GR/B3/1997.

## References

Boldea, M., Doroga, A., Dumitrescu, T. & Pescaru, M. (1996). Preliminaries to a Romanian Speech Database. In Proceedings ICSLP'96 (pp. 1934--1937). Philadelphia, PA, USA.

EAGLES Spoken Language Working Group (1994). Spoken Language Systems. Document EAG-SLWG-IR.2.

Chan, D., Fourcin, A., Gibbon, D., et al. (1995). EUROM - A Spoken Language Resource for the EU. In Proceedings EUROSPEECH'95 (pp. 867--870). Madrid, Spain.

Schmidt, M.S. & Watson, G.S. (1991). The evaluation and optimization of automatic speech segmentation. In Proceedings of EUROSPEECH'91 (pp. 701--704). Genova, Italy.

Zeiliger, J. & Serignat, J.F. (1991). EUROPEC Software v. 4.1 User's Guide. Institute de la Communication Parle, Grenoble, France.

Barry, W.J. & Fourcin, A.J. (1992). Levels of labeling. Computer Speech and Language, 6(1), 1--14.

Eisen, B. (1993). Reliability of speech segmentation and labeling at different levels of transcription. In Proceedings EUROSPEECH'93 (pp. 673--676). Berlin, Germany.

Garofolo, J.S., Lamel, L. F., Fisher, W.M., et al. (1993). DARPA TIMIT Acoustic-phonetic Continuous Speech Corpus. U.S. Department of Commerce.

Zue, V.W. & Seneff, S. (1988). Transcription and Alignment of the TIMIT Database. In Proceedings Second Symposium on Advanced Man-Machine Interface through Spoken Language. Oahu, Hawaii. Reprinted in (Garofolo et al., 1993), pp. 36--45.

Ljolie, A. & Riley, M.D. (1991). Automatic Segmentation and Labeling of Speech. In Proceedings ICASSP'91 (pp. 473--476). Toronto, Canada.

Brugnara, F., Falavigna, D. & Omologo, M. (1993). Automatic segmentation and labeling of speech based on Hidden Markov Models. Speech Communication, 12(4), 357--370.

Kipp, A., Wesenick, M.B. & Schiel, F. (1996). Automatic Detection and Segmentation of Pronunciation Variants in German Speech Corpora. In Proceedings ICSLP'96 (pp. 106--109). Philadelphia, PA, USA.

Wells, J.C. (1995). Computer-coding the IPA: a proposed extension of SAMPA. Department of Phonetics, University College, London, UK.

Gauvin, J.L. & Lamel, L.F. (1992). Speaker-Independent Phone Recognition Using BREF. In Proceedings 1992 DARPA Speech and Natural Language Workshop.

Rabiner, L.R. & B.H Juang,. (1993). Fundamentals of Speech Recognition. Englewood Cliffs, NJ:Prentice Hall.