

LABELING A ROMANIAN SPEECH DATABASE

Marian Boldea, Cosmin Munteanu

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

ABSTRACT

Speech databases are essential for acquisition of linguistic knowledge and speech technology developments, and both can be facilitated if the collected signal is accompanied by some form of annotation. Following the design and collection of a Romanian speech database including more than 10 hours of speech from 100 speakers, we are now working to label the signal files. This is done at a broad phonetic level, in a semiautomatic manner, including three steps: manual transcription of the signal files, automatic alignment of the phoneme labels in the transcriptions with the signals, and manual verification and correction of the aligned labels.

1. INTRODUCTION

As a first step towards automatic continuous speech recognition in Romanian [1], a speech database was collected [2], including both read and semispontaneous speech from 100 speakers, and is now labeled to increase its utility for linguistic and speech technology research. Following a brief review of its design and collection in section 2, this paper proceeds to detail the labeling process: theoretical considerations are introduced in section 3, the next section describes the database labeling proper, and the present status is given in the last.

2. DATABASE DESIGN AND COLLECTION

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 [3], and we chose as the primary application the development of speaker independent continuous speech recognition in Romanian, which implies a large speaker population. Another desirable characteristic would be a controlled number of occurrences for each of a predefined set of acoustic modeling units so that reliable model parameters could be estimated, and this implies

using specially designed text prompts to be read by the speakers.

In our case, as this is the first Romanian database of its kind, and there is no previous (to our knowledge) experience with modeling units for continuous speech recognition in Romanian, we decided that the database should allow for acoustic modeling at the phoneme level, and together with an as uniform as possible coverage of age and sex groups, this criterion was used to design the text prompts.

These were built around adapted translations from English of the passages used to collect the EUROM-1 database [4], grouped in 10 clusters through a heuristic procedure seeking a uniform a priori number of occurrences for each phoneme across clusters, and extended with a few (2-3) additional filler sentences, manually designed using an ad hoc editor, to raise to a minimum level the expected frequency of the least represented phonemes. 100 speakers (50 men, 50 women) were recorded, evenly distributed across five age groups (under 20, 20-29, 30-39, 40-49, 50 and over), each extended cluster being read once by one speaker from every sex and age group. To increase the 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.

Besides speaker independent continuous speech recognition development, another domain in which our database could be used is that of Romanian acoustic phonetics and phonology studies, whose lack we often felt unpleasantly during the database design stage, and for this purpose labeling the database was set as another objective. Speech signal labeling is though a very time consuming activity, and to make it easier using automatic procedures, we included four phonetically rich sentences, the same for all the speakers, to be labeled by hand and used as initialization material for phoneme models.

Other read materials similar to those in EUROM-1 were recorded to facilitate performance and diagnostic evaluation of speech recognizers (numbers, CVC words, CVC words in context), and extra read and semispontaneous materials were

collected for developing specific speech recognition applications (spoken and spelled names; telephone

The recordings, made in a sound proof room using the SAM protocols [3], with 20 KHz sampling rate, were all controlled for quality parameters (bias, clipping, SNR, noise components), and three CD-ROMs were produced.

3. THEORETICAL ASPECTS

3.1. Labeling level

Speech signal labeling means to temporally define discrete or overlapping parts of it and name them using symbols from an appropriately defined set corresponding to acoustic, physiological, phonetic or higher level linguistics terms [5], so it is an abstraction dependent on particular analytical and theoretical points of view, which in turn lead to various levels of labeling, each possibly encompassing different tiers in which separable aspects of the same level can be represented.

For speech technology applications, segmental and prosodic labeling are both important, and the most usual segmental levels, relevant to our work, are:

- physical, covering any labels defined exclusively with reference to physically defined events in an utterance, and which most clearly needs to be separated into different tiers corresponding to particular data acquisition or analysis procedures;
- acoustic-phonetic, including labels that describe acoustically homogeneous events in the speech signal in established phonetic terms (closure, release, aspiration, fricative noise, voicing, nasalization, etc.), without any claim about their linguistic function or distinctiveness, or their relation to physical events, although ultimately arbitrary decisions about their boundaries and the symbols set often require and are facilitated by a knowledge of what the sound is in phonetic-phonological terms;
- narrow-phonetic, whose labels characterize the phonetic quality of speech sounds in terms of phonetic symbols such as IPA or computer-compatible equivalents, and for which labeler's perceptual impression is the ultimate arbiter, as is for boundary placement; the most difficult problem at this level is boundaries placement where a perceived sequence of sounds does not match a corresponding sequence of acoustic events;
- phonemic, whose labels represent the functionally distinctive sound units in the language in which the utterance was spoken; it is the mediator between the signal and the lexicon (called also citation-phonemic for this reason) and, although labeling practice has

numbers; letter and digit strings; addresses; dates; the Romanian alphabet.)

not been of this type, useful for speech recognition and phonology work;

- broad-phonetic, often called phonemic, uses speech sound symbols with phonemic status to indicate non-phonemic (continuous speech) phenomena such as reduction and assimilation, and is intermediate in its degree of abstraction between narrow-phonetic and (citation-)phonemic; being the most economical, in that it maximizes phonetic information with minimal symbol set complexity, it is most commonly used in speech database labeling; given the symbol set, common with that in citation phonemic level, it is also the most appropriate for speech recognizers training and assessment.

Of these five levels, the physical and the (citation) phonemic ones are of no interest to our work, and the broad-phonetic level was chosen not only due to the above considerations, but also because it offers the highest labeling reliability [6], 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.

3.2. Automatic labeling

Done by hand, speech signal labeling is extremely time-consuming, and also involves decisions which are highly subjective, and various approaches were used to do it (at least in part) automatically to increase speed and consistency.

The first large acoustic-phonetic, and probably most known, reference speech database, TIMIT [7], was labeled using a semiautomatic procedure [8] 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) established themselves as fundamental tools in speech technology, HMM-based automatic segmentation and alignment became dominant [9], [10], [11], [12], 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 exist to be used for network generation, we chosen a labeling methodology similar to that used for TIMIT, i.e. manual transcription, HMM-based automatic segmentation and labeling, and manual verification and correction.

4. DATABASE LABELING

4.1. Speech signal transcription

Although final label files will be generated in SAM format [3] with SAMPA symbols used for the labels [13], the symbol set used for transcriptions is composed exclusively of single lower and upper case ASCII characters, given in table 1 together with their SAMPA correspondents and example words.

ASCII	SAMPA	Example word(s)
i	i	s <u>i</u> (and)
I	C	az <u>i</u> (today)
e	e	d <u>e</u> ge <u>t</u> (finger)
l	l	î <u>n</u> (in), c <u>â</u> nd (when)
@	@	dac <u>a</u> (if)
a	a	l <u>a</u> c (lake)
u	u	n <u>u</u> (no)
o	o	c <u>o</u> t (elbow)
j	j	ie <u>r</u> i (yesterday)
E	e_X	de <u>a</u> l (hill)
w	w	no <u>u</u> (new)
O	o_X	co <u>a</u> te (elbows)
p	p	ca <u>p</u> (head)
b	b	be <u>r</u> e (beer)
t	t	ti <u>m</u> p (time)
d	d	do <u>p</u> (cork)
k	k	ca <u>m</u> era (room)
g	g	gl <u>u</u> ma (joke)
T	ts	ta <u>r</u> a (country)
C	tS	ce <u>r</u> (sky), ce <u>a</u> i (tee)
G	dZ	ge <u>m</u> (jam)
f	f	fa <u>t</u> a (girl)
v	v	vi <u>n</u> (wine)
s	s	sa <u>r</u> e (salt)
z	z	zbu <u>r</u> (flight)
S	S	sa <u>p</u> te (seven)
J	Z	jo <u>c</u> (game)
h	h	ha <u>r</u> ta (map)
m	m	mi <u>c</u> (small)
n	n	na <u>s</u> (nose)
l	l	la <u>p</u> te (milk)
r	r	ro <u>s</u> u (red)
_ (underscore)	...	silence

Table 1: ASCII and SAMPA label symbols

Signal transcription is done based on listening and visual examination of waveforms, and includes continuous speech phenomena: assimilations, elisions, epentheses, etc.

4.2. 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 iteratively training phoneme and silence HMMs and using them for a Viterbi resegmentation of the signal files [14] (figure 1).

The acoustic processing operates on frames 12.8 ms long, spaced at 5 ms, preemphasized and weighted by a Hamming window, from which a 26 component vector including 12 autocorrelation LPC filtered cepstral coefficients, log energy, and their first derivatives, is computed.

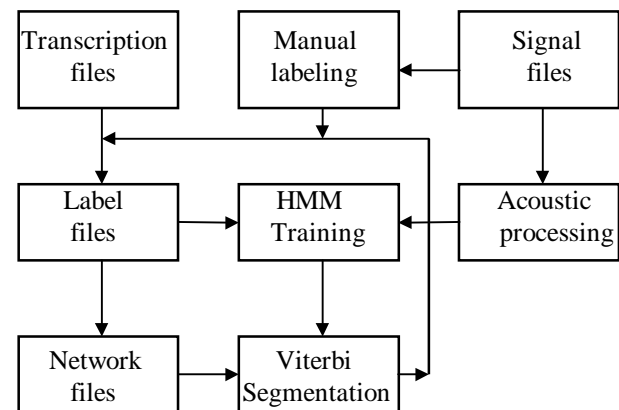


Figure 1: Automatic segmentation and alignment

Three-state left-to-right HMMs with Gaussian density mixture output probability functions are used, and to obtain a more rapid convergence of the training-resegmentation process, four phonetically rich sentences, the same for all speakers, were labeled manually and used to initialize the HMMs.

Output probability functions for every HMM state are initialized by using the manually labeled files to segment in three equal parts, corresponding to the three states, each occurrence of the corresponding unit, all feature vectors associated with a state being then grouped through an iterative K-means algorithm/Viterbi decoding procedure into a number of clusters equal to the number of Gaussians in each mixture; the resulting cluster means and variances are used as first approximations for Gaussian mean vectors and diagonal covariance matrices, and cluster weights - as density weights.

Next, a Baum-Welch procedure based on the same manually labeled signal files reestimates for each HMM both output probability functions and transition probabilities, the latter being initialized this way.

Finally, the training-resegmentation process uses all available signal and associated transcription files for concatenated HMM training [15], followed by

a Viterbi decoding of each signal file guided by a network generated from the associated label file. After a few iterations, the process converges, and the results of the last Viterbi decoding are saved as automatically aligned label files.

4.3. Label verification and correction

Once automatically generated label files are available, they are verified and corrected using a simultaneous synchronized display of signal waveform, spectrogram, and labels, based on listening and visual examination. To ensure consistency, a display at a constant 1 ms/pixel resolution is used, and rules are defined to be used in cases where different boundary placements are arguable.

5. PRESENT STATUS

The automatic alignment procedure was set up using about 200 sentences, plus the four phonetically rich sentences, read by a single speaker, and two-Gaussian mixtures proved to be sufficient in this case for obtaining good label-signal alignments. Quantitative results are not yet available, as the verification and correction procedure for this data set is still in progress.

Signal files from all speakers were transcribed, and phonetically rich sentences are now labeled by hand, to be used for separate (male/female) automatic alignments.

6. ACKNOWLEDGMENTS

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

7. REFERENCES

1. Boldea, M., Doroga, A., "Towards Automatic Recognition of Continuous Speech in Romanian", submitted to ROSE'97
2. Boldea, M., Doroga, A., Dumitrescu, T., Pescaru, M., "Preliminaries to a Romanian Speech Database", in *Proceedings Fourth International Conference on Spoken Language Processing*, vol. 3, pp. 1934-1937, Philadelphia, 1996
3. EAGLES Spoken Language Working Group, "Spoken Language Systems", document EAG-SLWG-IR.2, October 1994
4. Chan, D., Fourcin, A., et al., "EUROM - A Spoken Language Resource for the EU", in *Proceedings EUROSPEECH'95*, vol. 1, pp. 867-870
5. Barry, W.J., Fourcin, A.J., "Levels of labeling", *Computer Speech and Language* (1992) 6, pp. 1-14
6. Eisen, B., "Reliability of speech segmentation and labeling at different levels of transcription", in *Proceedings EUROSPEECH'93*, vol. 1, pp. 673-676
7. Garofolo, J.S., Lamel, L. F., et al., "DARPA TIMIT Acoustic-phonetic Continuous Speech Corpus", U.S. Department of Commerce, 1993
8. Zue, V.W., Seneff, S., "Transcription and Alignment of the TIMIT Database", in *Proceedings Second Symposium on Advanced Man-Machine Interface through Spoken Language*, Oahu, Hawaii, November 1988
9. Ljolie, A., Riley, M.D., "Automatic Segmentation and Labeling of Speech", *Proceedings ICASSP'91*, vol. 1, pp. 473-476
10. Brugnara, F., Falavigna, D., Omologo, M., "Automatic segmentation and labeling of speech based on Hidden Markov Models", *Speech Communication* (1993) 12, pp. 357-370
11. Kipp, A., Wesenick, M.B., Schiel, F., "Automatic Detection and Segmentation of Pronunciation Variants in German Speech Corpora", in *Proceedings Fourth International Conference on Spoken Language Processing*, vol. 1, pp. 106-109, Philadelphia, 1996
12. Wightman, C.W., Talkin, D.T., "The Aligner: Text-to-Speech Alignment Using Markov Models", in Van Santen, J.P.H., Sproat, R.W., Olive, J.P., Hirschberg, J., (editors), *Progress in Speech Synthesis*, Springer Verlag, New York, 1997
13. Wells, J.C., "Computer-coding the IPA: a proposed extension of SAMPA", Department of Phonetics, University College, London, 1995
14. Gauvin, J.L., Lamel, L.F., "Speaker-Independent Phone Recognition Using BREF", in *Proceedings 1992 DARPA Speech and Natural Language Workshop*
15. Lee, K.F., "Large-Vocabulary Speaker-Independent Continuous Speech Recognition: The SPHINX System", PhD Thesis, Carnegie Mellon University, Computer Science Department, April 1988