Segmentation and Labelling of Slovenian Diphone Inventories*

Jerneja Gros, Ivo Ipšić, Simon Dobrišek, France Mihelič, Nikola Pavešić
Faculty of Electrical Engineering
University of Ljubljana
Tržaška cesta 25
SI–1000 Ljubljana
Slovenia
jerneja.gros@fe.uni-lj.si

Abstract

Preparation, recording, segmentation and pitch labelling of Slovenian diphone inventories are described. A special user friendly interface package was developed in order to facilitate these operations. As acquisition of a labelled diphone inventory or adaptation of a speech synthesis system to synthesise further voices is manually intensive, an automatic procedure is required. A speech recogniser, based on Hidden Markov Models in forced segmentation mode is used to outline phone boundaries within spoken logatoms. A statistical evaluation of manual and automatic segmentation discrepancies is performed so as to estimate the reliability of automatically derived labels. Finally, diphone boundaries are determined and pitch markers are assigned to voiced sections of the speech signal.

1 Introduction

For the Slovenian language, several attempts were made in the past, where different aspects of a Slovenian text-to-speech synthesis (TTS) system were covered (Dobnikar95). Nevertheless, none of them succeeded in building a complete system, providing high quality synthetic speech. In the Laboratory of Artificial Perception, we started on text-to-speech synthesis one year ago (Gros96). Here we describe the acquisition of an appropriate diphone inventory in a first version of our Slovenian TTS system, which is supposed to serve as a reference system for future improvements.

We start with a brief overview of the different modules of the Slovenian TTS system, then we go on to describe how the existing diphone inventory was obtained. The acquisition of a labelled diphone inventory or adaptation of a speech synthesis system to synthesise new voices is manually intensive and prone to errors, therefore automatic procedures are required (Taylor91, Schmidt93, Cosi91, Ottesen93). In section 4 we explain how we intend to automatically derive additional diphone inventories for building new synthetic voices.

2 Slovenian TTS system

The different phases of the text-to-speech transformation are performed by separate independent modules, operating sequentially, as shown in Figure 1. Thus input text is gradually transformed into its spoken equivalent.

Grapheme-to-phoneme transcription

First, abbreviations are expanded to form equivalent full words using a special list of lexical entries. A text pre-processor converts further special formats, like numbers or dates, into standard graphemic strings. Next, word pronunciation is derived, based on a user extensible pronunciation dictionary and letter-to-sound rules. The dictionary is supposed to cover the most frequent words in a given language and a second dictionary helps with pronouncing proper names.

Prosody generation

Prosody generation assigns the sequence of allophones with some of their prosodic parameters (pitch frequency, duration). First, words are syllabified by counting the number of their vowel clusters and *duration of syllables* is modelled according to the speaker's normal articulation rate, depending on the number of syllables within a word and on the word's position within a phrase. Then, *segmental prosodic parameters* are determined for each allophone on the basis of the accent position within a word and its type. Finally, the global intonation contour of a phrase is determined (Sorin87).

Diphone Concatenation

Once the appropriate phonetic symbols and prosody markers are determined, the final step is to produce audible speech by assembling elemental speech units, computing pitch and duration contours, and synthesising the speech waveform. A concatenative TD-PSOLA diphone synthesis technique was used, allowing high-quality pitch and duration transformations directly on the waveform (Moulines90).

*This work was partly funded by the Commission of the European Community under COP-94 contract No 01634 (SQEL)

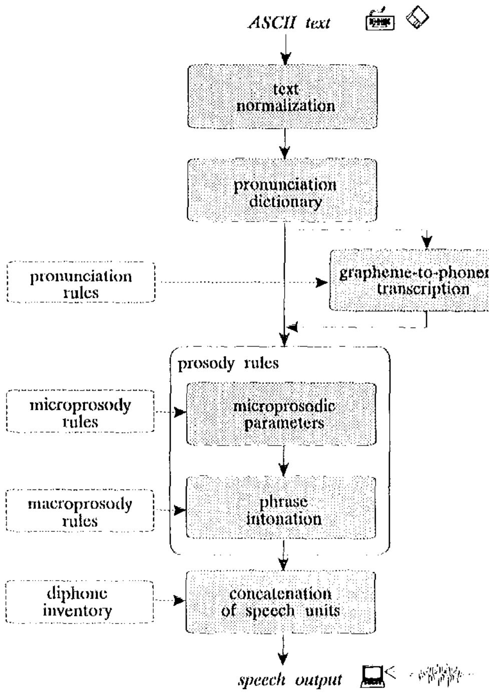

Figure 1: *Slovenian text-to-speech system architecture.*

3 Slovenian Diphone Inventory

In concatenation systems, both the choice and the proper segmentation of the units to be concatenated play a key role. Acoustic differences between stored and requested segments, as well as acoustic discontinuities at the boundaries between adjacent segments have to be minimised. Diphone units are most commonly adopted as a compromise between the size of the unit inventory and the quality of synthetic speech. A diphone is, generally speaking, a unit which starts in the middle of one phone, passes through the transition to the next phone and ends in the middle of this next phone. So the transition between two phones is encapsulated and does not need to be calculated.

Yet it is not clear whether speech segments should be extracted from nonsense plurisyllabic words, called logatoms, existing isolated words or meaningful sentences. Even the question of a best positioning of the units within the spoken corpus is still widely debated. Stressed syllables are longer, thus less submitted to coarticulation, which results in easily chainable units; while unstressed ones are more numerous in natural speech, so that producing them efficiently would both increase segmental quality and reduce memory requirements. Likewise, coarticulations are strongly subject to speaker's fluency, so that imposing a slow speaking rate results in more intelligible units. To a large extent, these issues are part of a necessary tradeoff between intelligibility and naturalness.

One diphone for every allophone combination possible in a given language is required. A Slovenian diphone inventory comprising 955 pitch-labelled diphones was created. In order to guarantee optimal synthesis quality, a neutral phonetic context in which the diphones needed to be located, was specified. Unfavourable positions, like inside stressed syllables or in over-articulated contexts, were excluded. The diphones were placed in the middle of logatoms, pronounced with a steady intonation. The exception is in the case where the silence phone is part of the required pair: there the diphone was word initial or word final. Speech signals were recorded by a close talking microphone using a sampling rate of 16 kHz and 16 bit linear A/D conversion.

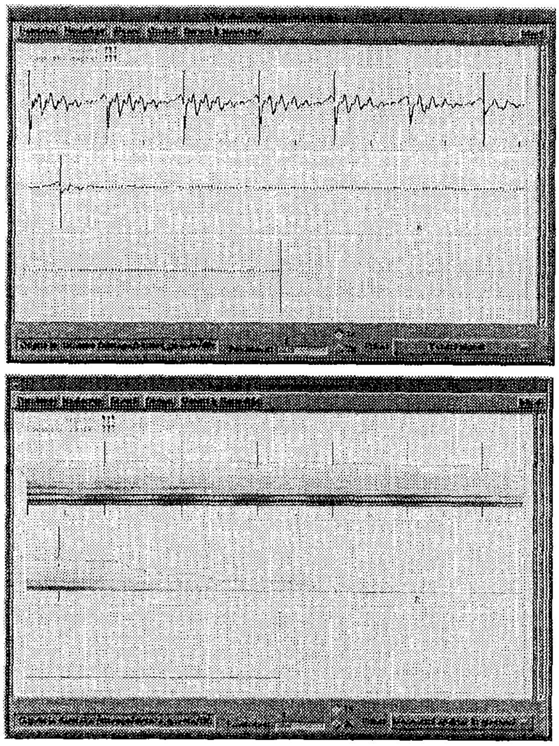

Figure 2: *Waveform* (above) *and spectral* (below) *representation of the diphone* ac. *Markers* L *and* R *are set at the pitch periods of the left part of the diphone and of the right part, respectively.*

After the recording phase, logatoms were hand-segmented and the center of the transition between the phones was marked, using information from both temporal and spectral representation of the speech signal. A special user-friendly interface was developed for this purpose, allowing editing, scaling, viewing, labelling and pitch-marking of the speech signal. First the approximate neighbourhood of a diphone was determined, then a fine labelling of its boundaries was performed and the center of the phoneme transition was marked. Finally, pitch markers were manually set for voiced parts of the corresponding speech signal. Figure 2 gives an example of the diphone *am* along with its spectrum.

To phonetically transcribe the logatom words we

phone	model
vowels	
i	(I, I, I)
é	(e, e, e)
ê	(E, E, E)
ə	(3, 3, 3)
ô	(O, O, O)
ó	(o, o, o)
u	(u, u, u)

phone	model
sonorants	
m	M
n	N
v-1	V
v-2	W
j	J
l	L
r	R

phone	model
nonsonorants	
p	(_, P)
t	(_, T)
k	(_, K)
b	(=, B)
d	(=, D)
g	(=, G)
f	F
h	H
s	S
š	Š
z	Z
ž	Ž
c	(_, C, S)
č	(_, Č, Š)

phone	model
silence	
silence	-

Table 1: *List of phones and their corresponding submodels used for Slovenian logatom segmentation. Symbol = represents a voiced closure while symbol _ represents an unvoiced closure.*

used a set of 34 symbols for allophones, which we adapted to the SAMPA standard requirements (Fourcin89)[1].

While concatenating diphones into words it suddenly turned out that there was a large discrepancy between the duration of allophones, as suggested by the prosody module, and the actual corresponding diphone duration stored in the diphone inventory. This happened due to the exaggerated eagerness of the speaker trying to pronounce the meaningless logatoms in a correct and clear way. Consequently, the quality of the synthetic speech was considerably affected and we are therefore planning to record another diphone inventory. As the transformation range for prosodic speech parameters needed for synthesising naturally sounding speech is large, the recording should thus be carefully controlled to achieve medium pitch and duration values.

4 Automatic Diphone Segmentation

Automatic speech segmentation procedures are powerful tools for including new synthetic voices and for updating and supplementing existing diphone libraries whereas manual diphone segmentation is a tedious, time consuming task, prone to errors. Therefore, in order to be able to synthesise speech in a variety of different voices, we decided to use procedures for automatic segmentation and pitch marking of spoken logatoms.

The extraction of diphones from the recorded words is performed in two stages. The first stage is the phoneme segmentation of logatoms, yielding a start point, transition center and end point for each phone. The second part of the diphone extraction procedure is to find the concatenation point of each phone.

Finally, pitch markers are to be determined for voiced parts of the signal. We intend to apply the SRPD (Super Resolution Pitch Determination) algorithm as it allows precise pitch determination (Medan91).

Hidden Markov Model Phone Segmentation

To solve the segmentation problem, methods for stochastic modelling of speech are used. Hidden Markov Models (HMMs) are stochastic finite-state automata that consist of a finite number of states, modelling the temporal structure of speech, and a probabilistic function for each of the states, modelling the emission and observation of acoustic feature vectors (Rabiner89).

To perform logatom segmentation we used the *Isadora* system, developed at the University Erlangen-Nuremberg (Schukat92). The *Isadora* system is a tool used for modelling of one dimensional patterns, like speech. It consists of modules for speech signal feature extraction, hard or soft vector quantization and beam-search driven Viterbi training and recognition. The *Isadora* system builds a large network of nodes that correspond to different speech events like phones, phonemes, words or sentences. The nodes are provided with a dedicated HMM in order to acoustically represent the corresponding speech event.

For system training, approximately half an hour of continuous speech recorded from a single speaker is required along with its orthographic transcription. The acoustical analyser delivers every milisecond a set of Mel frequency cepstral coefficients along with their slopes plus the energy of each frame. A phone level description is obtained using the orthographic transcription and a pronunciation dictionary. In the initialisation step the feature vectors are classified into 64 classes using a soft vector quantization technique. Using a phonetically labelled vocabulary a Baum-Welch training procedure is applied, and parameters of mono-

[1] A list of Slovenian SAMPA symbols together with their audio samples is available on the WWW on the address "http://luz.fer.uni-lj.si/english/SQEL/sampa-eng.html".

Phoneme	*Manual segmentation*			*Automatic segmentation*			Number of samples
	$\overline{x}$ [ms]	σ [ms]	confidence interval [ms]	$\overline{x}$ [ms]	σ [ms]	confidence interval [ms]	
A	82.00	30.80	2.08	84.10	30.20	2.02	852
B	21.70	7.06	1.07	42.30	18.40	2.79	171
C	75.90	17.60	4.88	86.30	25.30	7.02	52
Č	60.80	14.90	4.69	69.50	19.50	6.14	41
D	24.50	11.30	1.38	32.60	16.80	2.05	260
E	67.10	27.80	2.33	76.50	26.80	2.25	545
e	83.80	20.60	3.51	92.40	25.00	4.27	136
3	61.20	16.80	4.66	87.70	18.90	5.22	52
F	88.30	16.10	5.44	79.20	32.70	11.01	36
G	20.80	7.54	1.6	28.90	10.70	2.52	73
H	82.40	40.20	9.36	66.60	56.80	13.2	74
I	62.20	23.30	1.75	78.70	27.70	2.08	679
J	41.40	17.80	2.05	35.00	24.80	2.86	293
K	45.80	22.50	3.06	61.00	25.60	3.49	210
L	47.50	14.90	1.4	43.90	19.10	1.79	437
M	60.90	18.60	2.43	51.90	24.90	3.24	231
N	44.00	18.70	1.88	36.70	18.40	1.84	383
O	65.80	27.80	2.59	77.40	27.90	2.61	442
o	97.80	24.50	4.34	107.00	28.40	5.04	125
P	24.50	5.17	0.71	39.40	21.70	2.96	209
R	44.30	11.70	1.14	39.40	17.10	1.67	404
S	77.60	27.70	3.57	63.20	30.00	3.87	234
Š	76.20	22.30	6.65	76.70	29.20	8.71	46
T	31.60	13.90	1.2	36.70	19.80	1.72	510
U	65.40	22.90	3.1	75.30	25.40	3.43	213
V	54.60	10.30	1.8	48.90	19.50	3.39	130
Z	58.20	15.10	2.58	49.70	18.30	3.11	136
Ž	63.40	11.50	3.84	52.40	16.90	5.62	37
=	38.90	16.00	1.41	27.10	11.10	0.98	495
_	43.70	15.90	0.97	25.90	11.30	0.69	1040

Table 2: *Average phoneme duration, confidence interval and standard deviation of the population for manual and automatic segmentation.*

phone models are obtained. By applying the Viterbi alignment procedure, the training logatoms are automatically labelled using our monophone inventory.

Due to the properties of the Slovenian language some phones are composed of several phone components, like the stop consonants *k,p,b,d,t* and the affricates *c* and *č*. Such phones are described by multiple submodels. Table 1 gives the Slovenian phones and their corresponding submodels as they are used for logatom segmentation.

A preliminary statistical evaluation of manual and automatic segmentation discrepancies was performed on a much larger speech database than the logatom inventory itself as proposed in (Schmidt93). 150 spoken sentences were extracted from the Slovenian speech corpus GOPOLIS (Dobrišek96) concerning airflight timetable inquiries in total duration of 25 minutes. Average duration, confidence interval and standard deviation of the population for both manual and automatic segmentation are presented in Table 2.

The discrepancies between manual and automatic segmentation are considerable. Most of the problems arise when detecting bursts of plosives as the automatic procedure tends to shorten their closures considerably. The situation improves when plosives are taken as a whole, closures and bursts together.

As a result, a fully automatic segmentation of speech segments is hardly conceivable in the context of concatenation synthesis. As most phonological units originate via phonological considerations rather than on acoustic grounds, isolating them requires a deep prior knowledge of their specific features. Unsupervised segmentation, i.e. segmentation on acoustic principles only, often results in segments and sub-segments boundaries being misplaced, or just missing, while undefined ones appear. However, it can be used as a segmentation outline, the refinement of which has to be performed by a human expert.

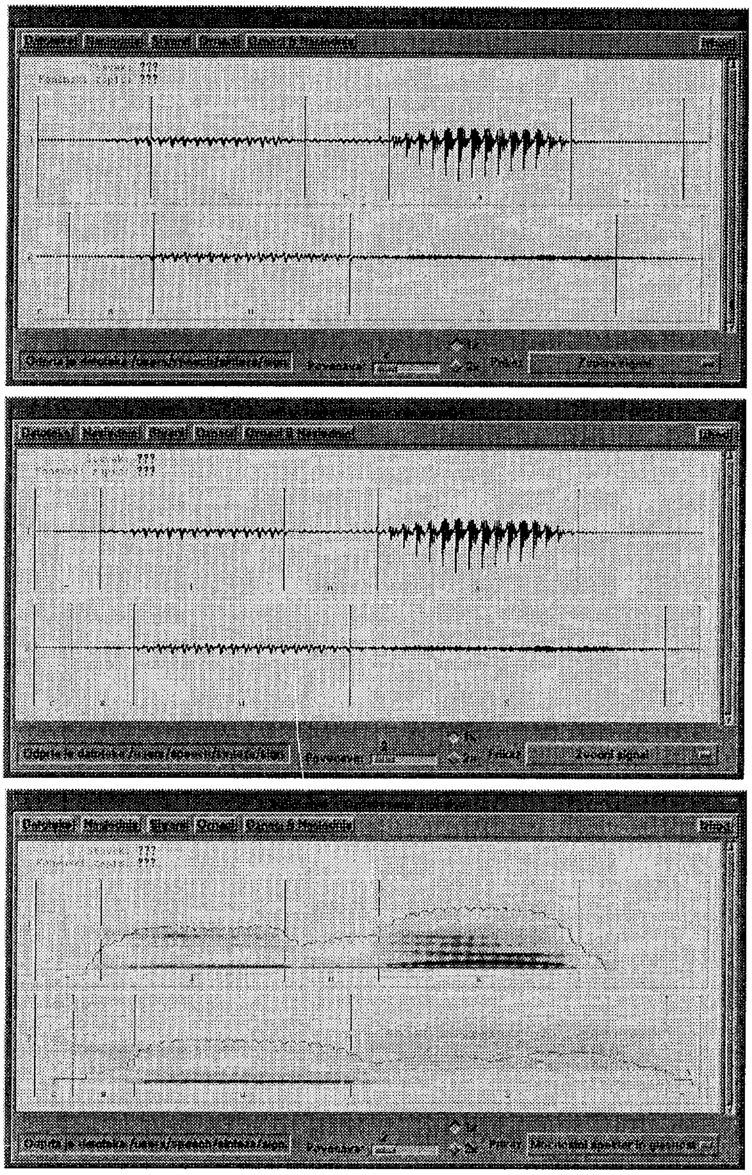

Figure 3: *Automatic* (above) *and manual* (center and below) *segmentation of the logatom* inacuš.

Thus automatic procedures can speed up the segmentation process, but they are not likely to suppress manual corrections, at least for obtaining highest synthesis quality with a given corpus.

Diphone boundaries determination

As the concatenation point of the diphones corresponds to the *center* of the phone, it is somewhere in the steady region of the phone. By studying the distances from the signal to the target values, (Ottesen91) claims that minimal distances tend to be just before the middle of the phoneme. We decided to divide each phoneme duration in a fixed ratio, 40 and 60%. Plosives are exception to this rule: they are divided just in front of the opening burst. A diphone boundary detection algorithm, minimising spectral discontinuities at concatenation points (Taylor91) may be investigated.

5 Conclusion

Diphone inventory acquisition for the Slovenian language was discussed. In order to avoid the tedious time consuming manual segmentation of logatoms, an automatic procedure, based on HMM models is considered. Thus diphone sets for new synthetic voices are easier to produce. Results of the statistical evaluation of manual and automatic segmentation discrepancies are given.

We expect the whole process of creating a new voice to be semi-automatic (with manual correction of stop-consonant boundaries), allowing the synthesiser to be retrained on a new voice in less than 3 days.

Acknowledgement

The authors wish to thank Tomaž Erjavec for proofreading of the text and his useful comments on the article.

References

A. Dobnikar and J. Bakran. 1995. A new approach for Slovene text-to-speech synthesis. *Proceedings Mipro95*. Opatija, Croatia. 265-268.

J. Gros et al. 1996. A text-to-speech system for the Slovenian language. *Eusipco96*. Trieste, Italy. Accepted for presentation.

P. A. Taylor and S. D. Isard. 1991. Automatic diphone segmentation. *Proceedings Eurospeech91*. Genova, Italy. 709-711.

M. S. Schmidt and G. S. Watson. 1993. The evaluation and optimization of automatic speech segmentation. *Proceedings Eurospeech93* Berlin, Germany. 701-704.

P. Cosi et al. 1991. A preliminary statistical evaluation of manual and automatic segmentation discrepancies. *Proceedings Eurospeech91* Genova, Italy. 693-696.

C. Sorin et al. 1987. A Rhythm-Based Prosodic Parser for Text-to-Speech Systems in French. *Proceedings XIth ICPhS*. Tallin, Estonia. 125-128.

E. Moulines and F. Charpentier. 1990. Pitch - Synchronous Waveform Processing Techniques for Text-to-Speech Synthesis Using Diphones. *Speech Communication*. 9:453-467.

Y. Medan et al. 1991. Super resolution pitch determination of speech signals. *IEEE Transactions on Signal Processing*. 39(1):40-48.

E. G. Schukat-Talamazzini et al. 1992. Acoustic modelling of subword units in the ISADORA speech recognizer. *Proceedings ICASSP92*. San Francisco, USA. 577-580.

L. Rabiner. 1989. A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. *Proceedings of the IEEE*. 77(2):257-289.

G. E. Ottesen. 1991. An automatic diphone segmentation system. *Proceedings Eurospeech93*. Berlin, Germany. 713-716.

A. Fourcin et al. 1989. *Speech Input and Output Assessment: Multilingual Methods and Standards*. Ellis Horwood Limited, John Wiley & Sons, New York - Chichester - Brisbane - Toronto.