LUKAS - a preliminary report on a new Swedish speech synthesis

Filipsson, Marcus; Bruce, Gösta

1997

Lund University, Dept. of Linguistics
Working Papers 46 (1997), 45–56

LUKAS – a preliminary report on a new Swedish speech synthesis

Marcus Filipsson and Gösta Bruce

Introduction

Background

For several years we have used speech synthesis at the Department of Linguistics and Phonetics, both as a research and as a teaching tool. The primary system has been the Swedish rule-based synthesis from Telia Promotor (formerly Infovox) and its developer's interface Rulsys from KTH. We have also used different implementations of Klatt synthesis on VAX and Macintosh. In the last few years, some successful projects (Bruce et al. 1995, Horne & Filipsson 1996a) have also used resynthesis of speech as a test method. In these cases, the recorded speech has been prosodically altered, and then resynthesized with an implementation of the PSOLA algorithm (Möhler & Dogil 1995).

Several previous projects at the department have aimed at modelling intonation in speech. The results of this research can easily be applied to speech synthesis. Merle Horne has demonstrated the need for tracking New/Given information in the text input to a speech synthesis system (Horne et al. 1993) and has also worked extensively on parsing text input prosodically (Horne & Filipsson 1996b), based on a Swedish dictionary which also contains phonetic transcriptions (Hedelin et al. 1987). Furthermore, the ProZodiag project led by Gösta Bruce and Björn Granström has been successful in generating a natural sounding intonation contour based on a limited set of prosodic labels (Bruce et al. 1995, Bruce et al. 1997). With these models and tools, we have many of the basic building blocks of a speech synthesis system. In late 1996 we arrived at the idea of trying to combine these building blocks into a speech synthesis system. Based on our experience with the PSOLA resynthesis, and given the existence of a Swedish rule-based synthesis, we decided to explore the possiblity of concatenative synthesis.

There are similar efforts being made at Telia Research AB (Gambäck et al. 1995).

The MBROLA project

At this point we made contact with the MBROLA team in Belgium (Dutoit et al. 1996). The idea of the MBROLA project is simple but very appealing: anybody can join as a developer, and record a diphone database of any language, segment it, and send it to the MBROLA team. They will process your database and return a compressed diphone synthesis database and an application which can run on many platforms and takes a string of phonemes as an input and produces speech. At the same time, your diphone database becomes public material, easily accessible from their WWW page. Thus, the aim of the MBROLA project is to facilitate free concatenative speech synthesis for as many languages as possible for non-commercial use.

It is important to point out that the MBROLA application is a phoneme-to-sound converter, *not* a text-to-speech system. The input is a table of phonemes with associated durations, and, optionally, one or more F0 values at percentages of duration calculated from the starting point of the phoneme. To create a complete TTS system from the MBROLA system one has to develop several components. The most obvious of these is a grapheme-to-phoneme converter which takes a string of input text and converts it to a string of phonemes coded with the SAMPA phonetic alphabet, as used by the MBROLA system. Furthermore, a system for assigning correct durations is needed. Finally, a system for generating an appropriate intonation contour as a list of time and frequency values is needed. As indicated above, we already possessed several of these components, developed within different projects, which made the idea of developing a complete TTS system seem feasible.

The pilot project

Introduction

Before we moved ahead and recorded a complete diphone database of Swedish we wanted to try a subset of the language. This was suggested by the MBROLA representative and was obviously a good idea. Designing, recording, and segmenting a complete diphone database is a monumental task and not something one would want to perform without proper testing of the conditions. During this pilot project all aspects of the process were tested. First, a representative number of Swedish phonemes was selected. Second, suitable words containing these diphones were selected and put in a carrier sentence. Third, the test sentences were recorded by a speaker. Fourth, the recorded speech was transferred to the computer and segmented. Finally, the segmented database was processed by the MBROLA team to produce a

Table 1. The phonemes of the pilot project.

IPA	*SAMPA*
p	p
t	t
k	k
s	s
ʂ	S
˛	C
n̡	n
r	r
ɪ	I
ʏ	u0
ç	O
ɖ˘	{:
ˊ	@
ß	rs

concatenative synthesis system. Thus, the following aspects were tested: the selection of context words and carrier sentences, the speaker's voice quality, the recording environment, the segmentation principles, and the suitability of the MBROLA algorithm for the Swedish language.

Selection of phonemes

The machine readable phonetic alphabet used by the MBROLA project is the SAMPA alphabet. This alphabet suggests 46 Swedish phonemes. For the pilot project we wanted to select a subset of these. The subset should be large enough to contain several of the specific Swedish phonemes and also large enough to enable the construction of real words and real sentences. On the other hand, as more phonemes are selected, the number of diphones to be recorded grows rapidly, so we wanted to keep the subset as small as possible. We finally arrived at 14 phonemes (Table 1). These were selected to contain enough consonants to build the Swedish consonant clusters containing (/s/, /p/, /t/, /k/, /r/), and also the Swedish fricatives /ʂ/ and /˛/. Furthermore, we wanted some typical Swedish vowels, both short (/ɪ/, /ʏ/, /ç/, /ˊ/) and long (/ɖ˘/). Finally we wanted one nasal consonant (/n/), and one post-alveolar phoneme (/ß/). We checked that we could build a few hundred words with these, enough to evaluate the result. These 14 phonemes lead to 14·14 + 14 initial + 14 final = 224 diphones. Some of these diphones were considered phonotactically impossible, and finally 210 diphones were decided on to be recorded. The decision to exclude phonotactically impossible phoneme combinations proved too hasty, as it turned out later (see below).

Recording the material

When recording the diphones one consideration was that they should not be uttered in isolation. Instead a suitable context word had to be selected. In this word, the diphone should not be emphatically stressed, but on the other hand not overly reduced. This trade-off led to some consideration, but in the pilot project we did not take other possible problems into account: hiatus, nasalization, or any other coarticulation effects besides those induced by phonemes immediately adjacent. Furthermore, a carrier sentence had to be selected. In the pilot project we decided on *Jag sa X en gång TILL* 'I said X one more TIME', where *TILL* is supposed to be focused. For each diphone we decided on three different context words, yielding three example sentences. This lead to a total of 630 sentences.

Next we had to decide on a speaker. The first author had shown in an earlier project (Horne & Filipsson 1996a) that he could produce speech of level intonation and intensity, and with a fairly constant speech rate. He was therefore selected as the speaker. The sentences were read from paper and recorded on DAT. During the recordings, the speaker made a conscious effort to speak in the way described above, and since the recorded material only amounted to about 30 minutes, he was fairly successful.

Segmentation of the database

The recorded sentences were transferred to a Sun workstation using a DAT-Link. With the *ESPS/waves+* and the *xlabel* programs the diphones were spotted and segmented according to the MBROLA standard. This encompasses a label at the mid-point of the first phone, a second label at the internal border between the phones, and a third label at the mid-point of the second phone. This segmentation task is the most work-intensive part of the whole database creation process, and it must be performed with precision and consistency. Naturally, there will be many cases where the best placement of a border is not obvious. The mid-point of sonorants is usually not too difficult to spot, neither is the mid-point of fricatives. For plosives, however, decisions have to be made as to where to place the borders, and this decision must be maintained throughout the material. In our case, we had three realizations of each diphone, which proved valuable, since it was not uncommon for one or two of the pronunciations to be unsatisfactory. This had several reasons: one or both phones in the diphone were too reduced, or perhaps spoken in the wrong way with regards to speech rate, intensity or intonation. There were of course also a few examples of mispronunciation.

Table 2. An example of the input to the MBROLA application. In this case it is the word *skånsk* [skçnsk] ‘Scanian’.

_	40	0.00	91		
s	90				
k	100	19.35	115	100.00	115
O	107	75.77	137		
n	82	100.00	63		
s	123				
k	151	100.00	63		
_	66				

The segmented database was sent to the MBROLA team for processing and returned as a phoneme-to-sound converting application.

Results

To test the pilot synthesis system, an application was developed which takes an input string and produces synthetic speech. The first part of this application is the information tracker (Horne et al. 1993). This process looks up all the words in the dictionary, and returns the phonetic transcriptions, the word class, and a tag of New/Given for each word. The next stage is the intonation generator (Bruce et al. 1995). This involves an automatic placement of prosodic labels, based on the predictions for stress given by the first stage. Focus is placed at the last new content word in the phrase in this first version. Intonation is then assigned as a table of time and F0 values. The next stage assigns durations to each phoneme. This is based on duration studies of Swedish (Elert 1964, Strangert 1985). A preliminary set of rules then manipulates these default durations while taking into consideration quantity, stress, focus, stress clash, and final lengthening. The result is sent to the MBROLA application as a table of phonemes with duration and intonation values (Table 2).

This preliminary TTS system was named LUKAS, (*Lunds Universitets KonkAteneringsSyntes* ‘Lund University Concatenative Synthesis’). It has a 14 phoneme inventory, enabling a vocabulary of a few hundred words. The limited vocabulary leads to some semantically uncommon sentences, and does not enable us to construct any longer coherent stretches of text. The system is capable of producing sentences like:

Pontus Kristensson är kär.	‘Pontus Kristensson is in love.’
Sussie Pontusson är konstnär.	‘Sussie Pontusson is an artist.’
Christer Persson är skånsk skinnknutte.	‘Christer Persson is a Scanian biker.’

Nisse är putt och purken. ‘Nisse is grumpy and sulky.’

No formal testing of the system has been performed – only informal demonstrations under varying circumstances. Although some of the sentences may be judged as uncommon, intelligibility of the output seems to be generally high, judging from the reactions of listeners so far. These listeners include both researchers at our department and members of the general public.

In the present system, no reductions are implemented. Thus, a word like *är* ‘is’ is pronounced [Qˇr] and not [E] or [´]. Another example is *och* ‘and’ which is pronounced [çk] and not [ç]. Reduced forms like these, and others, would probably lead to more natural output. It is in this context that phonotactically impossible combinations of phonemes appear. For example, the phoneme combination /´´/ (two schwas) would be considered impossible in Swedish, since /´/ can only be final, not initial. However, the string *Nisse är* ‘Nisse is’ [nIs´ Qˇr] could well be reduced to [nIs´ ´], thus leading to the sequence /´´/. Observations like these would have consequences for the selection of diphones in the main project.

There were a couple of realizations of diphones in the pilot project which were unsatisfactory. This was probably due to individual variation on the part of the speaker when recording the material. One such example was the diphone /kn/, where a click-sound could be heard when the diphone was placed in a context, such as *skinnknutte* [»SIn«knPt´] ‘biker’. Another example was the diphone /çt/, where probably a slight lip rounding at the end of the diphone made it possible to perceive it as /çpt/, leading to a word like *nikotinist* [nIkçtI»nIst] ‘nicotinist’ sounding like **nikoptinist* *[nIkçptI»nIst]. We considered two imperfect diphones out of 210 to be an acceptable error rate, however.

Some listeners pointed out that the synthetic voice sounded ‘tired’ and ‘uninspired’. This was probably due to the speaker’s effort to maintain an even speech rate and a flat intonation when recording the material. Perhaps he was too successful in these attempts, thus yielding a too mechanical voice quality. Another explanation could be that the F0 values used in the intonation generation possibly were slightly lower than the speaker’s own natural voice would produce. This would of course be easy to change.

In summary, we found the pilot project to be successful. We concluded that the voice quality of the speaker and the recording and segmentation circumstances were good enough. We also concluded that our preliminary rules for duration, and our model of intonation were sufficient to create

intelligible synthetic speech. We decided to continue with the main project to cover all Swedish diphones.

The main project

Introduction

Because of the promising results of the pilot project, we wanted to keep as many variables constant as possible in the main project. The same speaker would be used, as would the same recording device, computer, and software.

Selection of phonemes

The pilot project encompassed 14 phonemes. For the main project, we had to decide how many phonemes we needed to create natural sounding Swedish synthetic speech. The basic SAMPA phonetic alphabet contains 46 Swedish phonemes. This set of phonemes is basically the same as the one used in our Swedish dictionary. We decided to use these, but we extended it with four allophonic variants, namely the aspirated versions of /p/, /t/, /k/ and /˜/. These would be transcribed /pʰ/, /tʰ/, /kʰ/ and /˜ʰ/ or with the SAMPA alphabet *p_h*, *t_h*, *k_h* and *rt_h*. Thus, 50 phonemes were decided on (Table 3). Certainly, many other allophonic variants of phonemes could be included. The more variants we had covered, the more natural sounding the synthetic speech would be. The drawback was that the material to record and segment would grow rapidly for each new variant which was included. With these 50 phonemes we would get 50·50 = 2500 diphones, plus 50 initial and 50 final occurrences, leading to 2600 diphones. In the main project we did not want to presuppose anything about phonotactically impossible diphones. Instead, we decided to record all combinations, even some very unnatural sounding ones like /˜˜/ (two supradental *t*'s). This sequence would probably be pronounced with some vowel sound between them. An alternative would be not to record these combinations, but instead realize them in the TTS system as the two phonemes with a short /˙/ between them.

Table 3. The phonemes used in the main project.

IPA	*SAMPA*	*IPA*	*SAMPA*	*IPA*	*SAMPA*
p	p	j	j	ç	O
b	b	i˘	i:	a	a
t	t	e˘	e:	Q˘	{:
d	d	E˘	E:	ø˘	9:
k	k	y˘	y:	Q	{
g	g	ʉ˘	}:	ø	9
f	f	O˘	2:	ˊ	@
v	v	u˘	u:	˘	rt
s	s	o˘	o:	Í	rd
ʂ	S	A˘	A:	˜	rn
h	h	I	I	ß	rs
˛	C	e	e	ˋ	rl
m̊	m	E	E	pʰ	p_h
n	n	Y	Y	tʰ	t_h
N	N	P	u0	kʰ	k_h
r	r	O	2	˘ʰ	rt_h
l	l	U	U		

Even though we already had material recorded for 210 of the diphones, we decided to re-record them. We wanted the whole database to be recorded at the same time, in order to exclude the possibility of voice variation over time.

Some reductions in the material could be made: we believed that the aspirated allophones of /p/, /t/, /k/ and /˘/ would have the same left context as the unaspirated versions. Also, we decided to record the aspirated versions only before true vowels. These, and some other considerations led to a final diphone set of 2300.

Recording the material

When deciding on context words and carrier sentences in the main project, there were further considerations. First, since the database consisted of 2300 diphones, we wanted to reduce the work-load for the speaker as much as possible. For this reason a shorter carrier sentence was selected, namely *Ställ X DÄR* ‘Put X THERE’, where *DÄR* is supposed to be focused. We wanted it to be as short as possible but still effective, since when the speaker is to utter several thousand sentences every word we could remove would be a relief.

The choice of a suitable carrier phrase which would be effective for all test words was of course difficult. Our concern was to choose context words before and after the test word, which contained vowels and consonants which were reasonably neutral from a coarticulatory point of view. We therefore

decided on an [E] vowel in [stEl˘] and [dE˘r], which in the idiolect of the present speaker does not contain any extreme vowel quality. The choice of alveolar consonants ([l] and [d] respectively) at the edge of the context words is more debatable from a coarticulatory point of view but does not seem to be worse than labial, velar, or other places of articulation, which would be phonotactically available.

We also decided to have only two sample words for each diphone, instead of three.

The major part of the work when constructing the recording material, was in the selection of suitable context words. The Swedish dictionary was used extensively to automatically extract words with diphones in the right context. This produced a good starting point. Nevertheless, a large effort was needed to manually examine all 4600 context words while taking into account that the diphone should be:

- non-initial
- non-final
- not emphatically stressed
- not overly reduced
- not occurring in a coarticulatory unfavorable position

Moreover, the Swedish dictionary gave us a fairly broad transcription which did not offer details of pronunciation which sometimes turned out to be critical for accepting/rejecting a certain test word. For example, the transcription did not contain information about aspiration of voiceless plosives.

We are not certain of how important all these considerations are for the final result, but we wanted to try to predict as many problem situations as possible.

When the material was completed, the recordings were made. Recording such a huge material proved to be no small task. While constructing a diphone database, it is considered important to make all recordings in as short a period of time as possible in order to protect the material from variation. However, the database was so huge, and the recording material so repetitive and tiresome, that the recordings had to be spaced out to spare the speaker. Our recordings were made over a period of four days during the same week, producing a recorded material of 350 minutes, almost six hours.

In the pilot project, the recorded material was about 30 minutes. This relative small material made it possible for the speaker to concentrate on his speech rate and intonation throughout the material, thus speaking with a flat, ‘robot-like’ voice. In the main project, this proved to be very difficult. The

speaker could not maintain concentration, and inevitably reverted to a more natural sounding speech, with more variation in speech rate, intonation and articulation. We do not yet know if this will have any consequences for the result. Maybe it will have a positive effect on the liveliness of the voice. It might also lead to too reduced diphones in some instances.

Segmentation of the database

The segmentation of the database is currently under way in our laboratory. This task is time consuming, and will probably take several weeks. At the time of writing, only half of the material is segmented, and so far no serious problems have been discovered. The problems encountered when segmenting the pilot database, such as plosives, seem to repeat themselves in the main database, leading to the same considerations.

Results

No results can be reported from the main project yet.

Conclusion

We have made a pilot study of the possibility of developing a new Swedish concatenative speech synthesis. In the pilot study, 14 phonemes were selected. These were recorded as 210 diphones in context words. After segmentation, the diphone database was transformed into a phoneme-to-sound application by the MBROLA team in Belgium. This application was combined with a dictionary look-up system, duration rules, and an intonation generating system to produce a preliminary TTS system. The output was informally evaluated, and judged as promising with regards to intelligibility and segmental quality. The system was named LUKAS, (*Lunds Universitets KonkAteneringsSyntes* 'Lund University Concatenative Synthesis'). Based on the results of this system, we have decided to move forward with a complete diphone database for Swedish. In this main project, 50 phonemes were decided to be sufficient to cover enough variation. Context words and carrier sentences were selected. The material was recorded. The material is currently being segmented. We plan to have the complete database finished by the summer of 1997.

References

Bruce, G., B. Granström, M. Filipsson, K. Gustafson, M. Horne, D. House, B. Lastow & P. Touati. 1995. 'Speech synthesis in spoken dialogue research'. *Proceedings Eurospeech '95* (Madrid) 2, 1169-72.

Bruce, G., B. Granström, K. Gustafson, M. Horne, D. House & P. Touati. 1997. 'On the analysis of prosody in interaction'. In Y. Sagisaka, N. Campbell & N. Higuchi (eds.), *Computing Prosody*, 43-59. New York: Springer.

Dutoit, T., V. Pagel, N. Pierret, F. Bataille & O. van der Wrecken. 1996. 'The MBROLA project: towards a set of high quality speech synthesizers free of use for non-commercial purposes'. *Proceedings ICSLP '96* (Philadelphia) 3, 1393-96.

Elert, C.-C. 1964. *Phonological studies of quantity in Swedish.* Uppsala: Almqvist & Wiksell.

Gambäck, B., M. Eineborg, M. Eriksson, B. Ekholm, B. Lyberg & T. Svensson. 1995. 'A language interface to a polyphone-based speech synthesizer'. *Proceedings Eurospeech '95* (Madrid) 2, 1219-22.

Hedelin, P., A. Jonsson & P. Lindblad. 1987. *Svenskt uttalslexikon: 3 ed.* Tech Report, Chalmers University of Technology.

Horne, M. & M. Filipsson. 1996a. 'Implementation and evaluation of a model for synthesis of Swedish intonation'. *Proceedings ICSLP '96* (Philadelphia) Vol. 3, 1848-51.

Horne, M. & M. Filipsson. 1996b. 'Computational extraction of lexico-grammatical information for generation of Swedish intonation'. In J. van Santen, R. Sproat, J. Olive & J. Hirschberg (eds.), *Progress in speech synthesis*, 443-57. New York: Springer.

Horne, M., M. Filipsson, M. Ljungqvist & A. Lindström. 1993. 'Referent tracking in restricted texts using a lemmatized lexicon: implications for generation of prosody'. *Proceedings Eurospeech '93* (Berlin) 3, 2011-14.

Möhler, G. & G. Dogil. 1995. 'Test environment for the two level model of Germanic prominence'. *Proceedings Eurospeech '95* (Madrid) 2, 1019-22.

Strangert, E. 1985. *Swedish speech rhythm in a cross-language perspective.* Umeå: Almqvist & Wiksell.