Institute of Phonetic Sciences,
University of Amsterdam,
Proceedings 24 (2001), 57-63.

SYNTHESISED SPEECH WITH UNIT SELECTION

Creating a restricted domain speech corpus for Dutch

Betina Simonsen, Esther Klabbers and Louis Pols*

*IPO, Center for User-System Interaction, Eindhoven, The Netherlands

Abstract

This paper describes how I (= first author) have investigated what kind of units and how many should be included in a restricted domain application for Dutch with the purpose of creating synthesized speech with excellent phonetic and prosodic coverage. The restricted domain concerns number sequences such as phone numbers, credit card numbers, postal codes, currency amounts, bank account numbers and dates. An overview of the most common pronunciations was obtained by investigating the prosodic patterns of these number sequences. We chose the 'NH' (North Holland) group from the Dutch Polyphone Database as reference because the speaker that we used for our recordings also was a speaker of this dialect. At the end 268 sound files were selected as the basis for our number generator. As a final perceptual test, the prosodic naturalness of naturally spoken number sequences were compared with concatenated sentences.

1 Introduction

From this investigation the hypothesis sprung that only three forms were necessary in order to be able to generate all necessary digits and number sequences, namely a *neutral form* usually with a neutral pitch, a *continuing form* usually with a rising pitch and a *terminator form* with a falling pitch. Within these three categories different version of each digit exist i.e. with or without accent. The hypothesis served as a basis for the different versions in which the numbers were recorded. An evaluation experiment was set up where sentences that had not been used in the essential recordings were compared to identical sentences that had been generated from segmented units from the original recordings. This experiment revealed that the quality of the prosody in the concatenated utterances that were used was judged as just as good as in the natural sentences.

During the last couple of years, the focus has shifted from diphone synthesis towards corpus-based speech synthesis e.g. *selection based synthesis* or simply unit selection (Balestri, Pachiotti, Quazza, Salza and Sanst, 1999). The inventory of units used in the concatenation has been expanded from the basic diphone scheme. There are a number of directions in which this has been done, for instance changing the size of the units, the classification of the units themselves, and the number of occurrences of each unit. The quality of the synthesized speech depends on the corpus itself. It has to be large enough to contain the appropriate units and the relevant prosodic variations must be identified in order to keep the manipulations to an absolute minimum. Usually the corpus contains a large number of such units spoken by one speaker if the domain is restricted (Portele, 1998) and sometimes several speakers if the domain is unrestricted.

Diphones are speech segments cut in the stable part of the phonemes, spanning two half-phonemes and including in the middle the phonetic transitions. Diphone synthesis systems usually maintain a high level of intelligibility and preserve some of the co-articulation, however the naturalness still leaves a great deal to be desired (Klabbers, 2000). This is mainly due to the manipulation of the duration and the F_0.

The unit-selection synthesis approach is based on the concatenation of pre-recorded units but manipulation is not obligatory. This is because the speech corpus should contain each synthesis unit in different prosodic settings (Balestri et al., 1999; Beutnagel et al., 1999; Stöber et al., 1999; Black and Taylor, 1999). If a word is not present in the database it can simply be built up from smaller units.

The difference with traditional diphone syntheses is that the selection algorithm attempts to select the largest possible units in the database, thereby minimizing the number of junctions. A cost function is evaluated for each unit combination and the sequence of units that has the lowest cost is then chosen. The results of this method are usually judged to be more natural sounding than diphone synthesis. Although the synthesis systems of today are often highly intelligible and prosodically natural, the output still lacks the variety of natural human speech. Campbell (1997) claims that this is because too much emphasis has been put on the intelligibility of the text rather than on the naturalness of the voice.

The research I (= first author) am reporting about here was done at IPO, Center for User System Interaction, at the Technical University of Eindhoven under the supervision of Esther Klabbers (IPO) and Louis Pols (UvA). This MA-thesis is a round off of my education at the University of Amsterdam, where I obtained a Master of Arts (M.A.) in Computational Linguistics with the specialization Speech Technology.

2 Establishing prosodic structure

There is not just one correct prosodic realization for a sentence. All sentences can be realized prosodically in numerous different ways without one being more correct or acceptable than the others. A person who is speaking fast tends to prosodically mark fewer boundaries, or mark them differently, than someone who is speaking at a lower speech rate. Intonation is also closely related to age, mood, style of speaking, interest in the subject, and relationship with the listener, etc. In short this means that every single person has his/her own 'sound'. This means that finding prosodic features that are common in certain situations and with relation to certain subjects is fairly difficult.

I collected the most common pronunciation patterns and forms in the following manner: Firstly I listened to the i.e. telephone numbers from the Polyphone corpus and looked at the prosodic pattern of each number. I compared the extracted pitch tiers of each number in ‘PRAAT’ (a system for doing phonetics by computer by Paul Boersma and David Weenink, 1992-1998). If one specific pronunciation pattern was used at least ten times by different speakers it was selected as being frequent and entered in a table as an independent form. The forms that were captured this way are thus visual representations of the pronunciations chosen by the speakers in the Polyphone Database. The visual representation of the number is largely (always) responsible for the chosen pronunciation. If the number is written in clusters of two then this is most likely the way it will be pronounced. When the number is written differently than in clusters of two, then the pronunciation changes. See Table 1 below for an example of a credit card number and a phone number.

Table 1. Prosodic and schematic representation of the pronunciation of a credit card and a phone number.

Type	Form	Prosodic structure	Schematic representation
0-9	46 28 04 54 56 25 49 68	“vier zes / “twee acht / “nul vier / “vijf vier / “vijf zes/ “twee vijf / “vier negen / “zes acht /////	[1] [1] [1] [1] [1] [1] [1] [1] [1] [1] [1] [1] [1] [1] [1] [1]
xxxx-xxxxxx	0318-513 391	“nul drie een “acht / “vijf dertien / “drie een en negentig /////	[1] [1] [1] [1] [1] [10] [1] [1] en [10]

[1] is a digit (0-9), [10] represents tens (10-14) and [100] represents hundreds. The “ indicates accent, / indicates a phrase boundary and ///// indicates a final phrase boundary.

After having captured the prosodic structures and accent patterns of the earlier mentioned numbers and digits and analyzed and compared these, I came to the conclusion that only three categories of all relevant numbers should be sufficient in order to create natural sounding synthesized speech: a *neutral form* usually with a neutral pitch, a *continuing form* usually with a rising pitch and a *terminator form* with a falling pitch.

3 Adaptation and implementation of the hypothesis

Firstly, I created a list of the different types of numbers that I had investigated. Then I listed the versions and realizations necessary within each of these categories i.e. neutral form with a neutral pitch, continuing form with a rising pitch and no accent, continuing form with a rising pitch and accent, terminator form with a falling pitch.

Secondly, I extracted the three earlier mentioned forms from this list. Within each form I listed which numbers or units were necessary for each type of number and whether the digits needed to be accented or not.

Thirdly, I produced a printed list of utterances with all digits that had to be pronounced using the right form and pitch manner i.e. with or without accent. I also created a test page with between 4 to 10 written examples of numbers from each of the 7 number types. The test page was recorded without giving the speaker any instructions as to pronunciation. I intended to use this as a cross-reference for later evaluation of the synthesized speech.

Last, I recorded all the utterances spoken by one female speaker at IPO in Eindhoven in a soundproof room. The material was recorded digitally on a Digital

Audiotape. The test page was recorded first of all. All the utterances were recorded twice.

3.1 Processing the speech material

The speech material was transferred from the Digital Audiotape to GIPOS (**G**raphical **I**nteractive **P**rocessing **O**f **S**peech, 1992-1998) which is a program created at IPO to view, play and manipulate waveforms, spectrograms and other forms of speech data. The speech material was recorded in files of a fixed duration of 5 minutes so as not to make the files too large.

The files were segmented according to the sentences that were originally recorded. This resulted in 2 * 268 smaller sound files. The individual sentences were phonetically transcribed by hand according to the Dutch SAMPA alphabet (**S**peech **A**ssessment **M**ethods **P**honetic **A**lphabet, 1987-1989), which is a computer alphabet, used to represent sounds through ASCII-symbols. We listened to all the sound files and the files that were judged the best (subjectively), by my supervisor at IPO and myself, were collected into one set of 268 sound files (*.aiff). Best in this respect meant the file that was most representative of the intended pronunciation form.

3.2 The naturalness evaluation experiment

In this evaluation experiment I wanted to test to what degree listeners would judge the naturalness of the prosody of the concatenated sentences. I randomly selected 3 sentences from each of the seven number categories of the test recordings and generated identical sentences by concatenating manually segmented units that were derived form the original recordings.

In order to uphold one of the principles of unit selection, which is to avoid modeling or manipulating the prosodic parameters of the units, I selected the units as large as possible and as close to each other as possible, whenever it was possible, from the sentences that had been recorded.

I chose 10 naïve listeners because in my opinion experienced listeners would probably listen more closely to the quality of the synthesis and would then rather differentiate between natural speech and re-synthesized speech than concentrate on the quality of the naturalness of the prosody. The listeners were given a written introduction prior to the test. They were informed that the sentences they would hear were all statements rather than say questions or otherwise. They then listened to one randomization of the 21 ‘natural’ sentences and the 21 manually generated synthesized sentences, 42 sentences in total. After each sentence the subjects were asked to judge and write down the naturalness of the prosody of the sentence on a scale form 1 to 5 on a separate form: 1 meaning that they totally agreed that the sentence sounded unnatural and 5 meaning that they thought the sentence to sound completely natural. The answers had to be given within 7 seconds.

I deliberately chose to conduct the test as a ‘Comparison Category Rating’ (Andersen, Dyhr, Engberg Nielsen, 1998) because I didn’t want to force the participants to make a choice between two sentences in each case. Listeners are very good at detecting synthetic speech and the point was to make them focus on the naturalness of the speech prosody rather than whether it was synthesized speech or natural speech. However, in my subsequent data analysis, I chose to pair the normal with the concatenated sentences in order to see how far apart they were in terms of perceived naturalness scores.

3.2 Statistical analysis

Normally when using statistical information in research it is customary to show that there is a – preferably – significant difference between the different categories. In this experiment we wanted to show that the difference between the two categories was not significant. Several non-parametric tests were carried out in order to get an overview of the scores and of the judgments of the listeners. These were done in SPSS version 10 for Windows (**S**tatistical **P**ackage for **S**ocial **S**ciences). Firstly, the homogeneity of the listeners was investigated, resulting in $\alpha = 0{,}8835$, which means that the listeners were very homogenous (reliable) in their judgments. Secondly, the Wilcoxon Test was carried out to investigate whether the differences in scores between the paired normal-concatenated sentences was significant or not. Thirdly, a Wilcoxon signed ranks test was carried out to look at the distribution of the scores within the sentences.

3.2 Results

The Wilcoxon test showed that in all but three cases the differences between the normal and the concatenated sentences were perceived by the listeners as being not significant (1: p=0.046, 2: p=0.010, 3: p=0.010).

The Wilcoxon signed ranks test showed that in 60 cases (28,57%) the concatenated sentences were perceived as sounding prosodically more natural or equally natural to the normal sentences. In 65 cases (30,95%) the normal sentences were perceived as sounding prosodically more natural than the concatenated sentences and in 85 cases (40,48%) they were judged as sounding equally normal. The normal sentences have a small preference (5 cases) but the 'normal=concatenated' judgments were still in the majority. The difference was still not significant (p=0.721) confirming the findings in the Wilcoxon test.

4 Conclusion

In this thesis the following question was investigated:

- Which units should be included in an ideal corpus for unit selection (units being digits and numbers) in order to be able to generate all necessary number sequences in Dutch?

In the evaluation experiment the listeners confirmed that three forms are sufficient in order to be able to produce natural sounding speech in the sense that they judged the concatenated sentences to be just as natural sounding as the natural sentences. The difference between the natural and the concatenated utterances was a mere 2,38% in favor of the natural sentences.

However, since this was a small-scale evaluation (21 test sentences only) these results cannot be seen as representative of how the selection will be done in the unrestricted domain corpus and consequently how the output will sound. The evaluation was only intended to test a priori whether the three forms (neutral form, terminator form, continuing form) are sufficient for the restricted domain that we have dealt with in this thesis.

5 Discussion

When I started working on this project and even after I had identified the three forms I was still not convinced by van Santen (1997) when he argued that even half a billion units will not be enough to produce satisfactory natural sounding synthesis for an unrestricted domain system. We only investigated a rather small, restricted domain and ended up with 268 utterances, which is not much. However, these utterances, when segmented, will result in many thousands of units. If one then keeps in mind one language and all the domains contained therein I now believe that he may be right. Van Santen (1997) also suggests that prosodic modeling is indeed necessary to achieve an acceptable degree of naturalness for unrestricted domain TTS and I must say that after having tried to concatenate all necessary units without manipulating any of the signal parameters I am almost convinced of that too now. If one does not wish to manipulate the speech units in any way one must be absolutely sure that all units necessary are contained in different prosodic settings or one will never be able to produce natural sounding synthesized speech. Optimally when selecting units from the database they must be within a close proximity of each other or the spectral and intonational differences will be too great. Either one will need to have an extremely large database or one will need an extremely good and detailed selection procedure. This means that the programming will become very extensive and complicated, thus possibly worsening the result of the unit selection process because it will be too selective.

But how will we know when we have all the units it takes to be able to generate unrestricted domain synthesised speech? Do we really possess enough knowledge about the languages and the structure of the languages and the choices we make when speaking our own language to be able to accomplish this and know what these units are? Why is it for instance that one person never uses anything else but single digits when pronouncing his credit card number whereas he/she does use tens when he pronounces his bank account number? What degree of recognition is necessary for someone to use tens instead of digits and is it at all related to degrees of recognition or familiarity? I fear that before we know enough about what motivates us to make certain decisions and on what basis, we won't know what an ideal corpus for a corpus-based synthesis should look like. In the earlier days the power of the computers and storage room were a key factor and limiter but nowadays this is less of a problem. In my opinion the most important limits today are the limits of the speaker and the creator/researcher.

References

Andersen O., Dyhr N.-J., Engberg I.S. & Nielsen C. (1998): "Synthesized short vowels from their long counterparts in a concatenative based text-to-speech system", Proc. 3rd ESCA/COCOSDA Workshop on Speech Synthesis, Jenolan Caves, Australia, pp. 165-170.

Balestri M., Pacchiotti A., Quazza S., Salza P-L. & Sandri S. (1999): "Choose the best to modify the least: A new generation concatenative synthesis system", Proc. Eurospeech'99, Budapest, Hungary, Vol. 5, pp. 2291-2294.

Beutnagel M., Mohri M. & Riley M. (1999): "Rapid Unit Selection from a Large Speech Corpus for Concatenative Speech Synthesis", Proc. Eurospeech'99, Budapest, Hungary, Vol. 2, pp. 607-610.

Black A.W. & Taylor P. (1997): "Automatically Clustering Similar Units for Unit Selection in Speech Synthesis", Proc. Eurospeech'97, Rhodes, Greece, Vol. 2, pp. 601-604.

Campbell N. (1997): "A step in the direction of synthesising natural-sounding speech", retrieved from the Internet: http://www.itl.atr.co.jp/chatr.

Gipos User Manual 1992-1998: From the web site of the IPO, Center for User-System Interaction: http://www.ipo.tue.nl/ipo/gipos/#A1.

Klabbers E. (2000): *Segmental and Prosodic Improvements to Speech Generation*, Chapter 2, "Generating High Quality Speech"., Ph.D. thesis, Technical University of Eindhoven, pp. 7-25.

Portele, T. (1998): "Just CONcatenation – A corpus-based approach and its limits", Proc. 3rd ESCA/ COCOSDA Workshop on Speech Synthesis, Jenolan Caves, Australia, pp. 61-71.

SAMPA (1987-1989): From the web site of The Department of Phonetics and Linguistics at University College London: http://www.phon.ucl.ac.uk/home/sampa/dutch.htm.

Santen van J. (1997): "Prosodic Modelling in Text-To-Speech Synthesis", Proc. Eurospeech'97, Rhodes, Greece, Vol. 1, pp. HN-19-28.

Stöber K., Portele K., Wagner P. & Hess W. (1999): "Synthesis by Word Concatenation", Proc. Eurospeech'99, Budapest, Hungary, Vol. 2, pp. 619-622.