Speech synthesis and electronic dictionaries

Arthur Dirksen

Fluency, Amsterdam

Abstract

Speech synthesis software can be used in electronic dictionaries to generate audible renditions of phonetic transcriptions, idioms and example sentences. This paper discusses advantages and disadvantages of using synthesized rather than recorded speech for these purposes, and examines one approach in some detail.

1 Introduction

Speech synthesis came out of the laboratory when sound cards, CD-ROMs, large hard disks and fast processors appeared on standard consumer PCs, sometime around 1995. These advances in technology also enabled electronic dictionaries and encyclopedias of (almost) unlimited size and scope. Since one of the purposes of a dictionary is to indicate how a word is pronounced, an electronic dictionary has an advantage over its paper equivalent in that it can do this using sound rather than symbols. The easiest and most straightforward way to add sound to a dictionary is to use recorded natural speech, and this is certainly the most common approach. However, speech synthesis provides a more general, flexible approach, which is well worth the additional investment.[1]

The most obvious advantage of using recorded human speech in an electronic dictionary is the naturalness and liveliness that can be obtained. But the newer speech synthesizers sound surprisingly natural, to the extent that it becomes difficult to distinguish synthesized from natural speech. Also, naturalness is but one aspect of speech. It is certainly the most important because it dominates other aspects: if speech sounds highly unnatural it will also be rather unintelligible, not just unpleasant to listen to. But once a sufficient level of naturalness is obtained, other aspects matter as well. In the context of an electronic dictionary, the user will want to know which sounds make up the pronunciation of a word, which syllable carries the main stress, how the distribution of stress influences the rhythm of the pronunciation (in stress-timed languages like English and Dutch), and not much else. If a speech synthesizer can do a convincing imitation of this, it is certainly good enough.

The advantages of using a speech synthesizer include the following:

- *scalability* With recorded speech, the amount of storage needed ultimately becomes problematic (and becomes problematic rather quickly on small devices). With synthetic speech, as each word is generated on the fly (using some kind of phonetic transcription as input), the same software can generate one million words just as easily as 10,000.

---

[1] On a more personal note, I came out of the laboratory when in 1996, after several years of academic research into linguistics, speech synthesis and text-to-speech, I started developing commercial text-to-speech software for Dutch, and founded Fluency (www.fluency.nl). Sieb Nooteboom helped in many ways with this "coming out," for which I am very grateful. Sieb was especially intrigued by the collaboration between Fluency and Van Dale Lexicografie on the topic of this paper, and always liked to mention that he had already discussed this possibility with Van Dale many years ago, when sound cards were still the size of a refrigerator. Which is why I think this paper is a suitable tribute to Sieb. In addition, I would like to thank Ludmila Menert, Josée Heemskerk, Johan Zuidema, Rik Schutz and Bram Wolthoorn for their various contributions to the work discussed here.

- *extensibility* New words can be added by adding a phonetic transcription (which will typically be done by the editors of the dictionary). There is no need to go back to the recording studio. Also, it is possible to include the pronunciation of not just head words, but all inflected forms of a word, or to offer audible pronunciation of idioms and example sentences. The latter is especially useful in a learner's dictionary, allowing the user to exercise the pronunciation of full sentences.
- *customizability* A speech synthesizer may allow the user to indicate preferences for a variety of aspects of the synthesized speech, such as the voice to be used (male, female), pitch and speech rate, or even speaking style (rather formal or somewhat informal). In addition, the dictionary publisher has the possibility to tune the pronunciation details to the intended audience of a given title (examples will be given in section 4).

A rather more subtle advantage, which need not be appreciated by users, is that a speech synthesizer establishes a formal and consistent relation between a phonetic transcription on the one hand, and an audio signal on the other. With recorded human speech, this can never be guaranteed, and the recording can only be regarded as an example of how a word is pronounced. With a speech synthesizer, to the extent that it operates correctly, what you see (in terms of phonetic symbols) is what you get (in terms of audio samples). For this reason, a dictionary publisher might wish to use speech synthesis to develop and evaluate a database of phonetic transcriptions.

This paper discusses collaborative research and development by Van Dale Lexicografie, a Dutch dictionary publisher, and Fluency, which produces text-to-speech software for Dutch. Section 2 briefly summarizes the development of a large pronunciation database for Dutch at Van Dale. Next, section 3 gives a rapid overview of the Fluency text-to-speech software, and its application in electronic dictionaries. Section 4 discusses the rule-based system that translates a phonetic transcription into a full specification for the MBROLA diphone synthesizer (Dutoit, 1997), which is used by the Fluency software for audio generation. These rules implement phonological and allophonic processes, and assign durations and pitch. Subtle (and admittedly non-modular) interactions between phonological and phonetic rules in our system define a formal relation between phonetic transcriptions and speech which is highly flexible and can easily be customized to suit a variety of applications. Section 5 indicates how example sentences can be synthesized correctly using annotations for pitch accent placement.

2 A database of phonetic transcriptions

The most economic way to produce dictionaries of varying size and scope for different audiences (children up to language professionals) is to derive them from a large database. Rather than updating each dictionary title for each new release, editors update the database directly, which saves not only time and money, but also improves consistency among the various titles. The construction and maintenance of such a database, however, is no small matter, especially if a dictionary publisher has a lot of legacy material. In the past decennium, Van Dale made a leap forward in this respect by constructing a database encoding the linguistic properties of over 1.5 million Dutch word forms. The properties encoded include spelling (as well as common misspellings), hyphenation and syllabification, part of speech, syntactic features, morphological structure, frequency of occurrence (in a large and frequently updated corpus of newspaper text), and, of course, pronunciation. In the construction of the database, language technology software and tricks-of-the-trade were heavily used, but everything was manually checked and edited. The database is used in-house to automatically enrich dictionaries, but it also serves as a source of data for language and

speech technology applications of Van Dale (Froon, den Hartog & Zuidema, 1999) and third parties.

The phonetic transcriptions in the database are not tied to a particular set of phonetic symbols such as IPA or SAMPA. Instead, they use a more abstract notation, from which actual transcriptions are derived: IPA for the professional dictionary, respelling for the children's dictionary, or whatever is needed for a particular application. Also, the master transcriptions include morphophonological information, such as secondary stresses, syllable and compound boundaries, and applications of phonological rules. Some examples:

- In a case such as *woordenboek* 'dictionary', a compound boundary is indicated in the transcription between *woorden* 'words' and *boek* 'book', and a secondary stress is indicated on the second part of the compound.
- In the transcription of the word *spinnenweb* 'spider web', it is indicated that the final consonant, which is pronounced [p] due to Final Devoicing, is underlying /b/ (which surfaces as [b] in the plural *spinnenwebben*).
- Although both *komen* 'come' and *gaan* 'go' end in /n/, the /n/ in the plural suffix *-en* may be deleted or reduced (unless /n/ is resyllabified as in *ze komen en gaan* 'they come and go'). Hence, in the transcription for *komen*, it is indicated that we are dealing with a deletable /n/.
- In *postbode* 'mail man', the /t/ may be deleted in casual speech, and this deletion would trigger Progressive Voice Assimilation: /s/ → [z]. Again, the transcription indicates the special nature of such consonant clusters.

All this extra information, which is not usually found in phonetic transcriptions for a dictionary, is a boon for applications and research in the area of language and speech technology. For example, in a lexicon for a morphophonological parser one will want to include underlying rather than surface forms. But also, the extra information allows the phonetic transcriptions to be tuned to a particular audience.

3 From text to speech

Conversion of text to speech is usually a three-stage process. In the initial stage, input text is assigned a phonetic transcription by looking up words in a lexicon, applying morphological or grapheme-to-phoneme rules for unknown words, as well as special rules for numbers, punctuation, dates, e-mail addresses, and so on. In the second stage, the phonetic transcription is modified by phonological rules (to the extent needed), each phoneme is assigned a contextually appropriate duration, and pitch targets are set to create a fitting intonation pattern. Finally, in the third stage the actual speech synthesis occurs, using one of several available methods (e.g., formant synthesis, diphone concatenation or unit selection).

Figure 1 gives a visual impression of how the Fluency text-to-speech software implements the three stages (for audio examples, go to www.fluency.nl). The second panel shows the phonetic transcription. The third panel displays the allophonic transcription (top), durations (bottom) and pitch targets (middle).

The software uses a lexicon of approximately 180,000 word forms. These were selected from the database described in section 2, using the lemma frequency of the word forms as the main criterion, with manual additions to optimize the lexicon for text-to-speech conversion. In order to cope with the fact that Dutch text is littered with English words and terminology, we also included some 15,000 English word forms, transcribed as well as one can expect using

the Dutch phoneme symbols. This lexicon covers arbitrary Dutch text rather well, and is augmented by compound analysis and grapheme-to-phoneme rules as backup mechanisms.

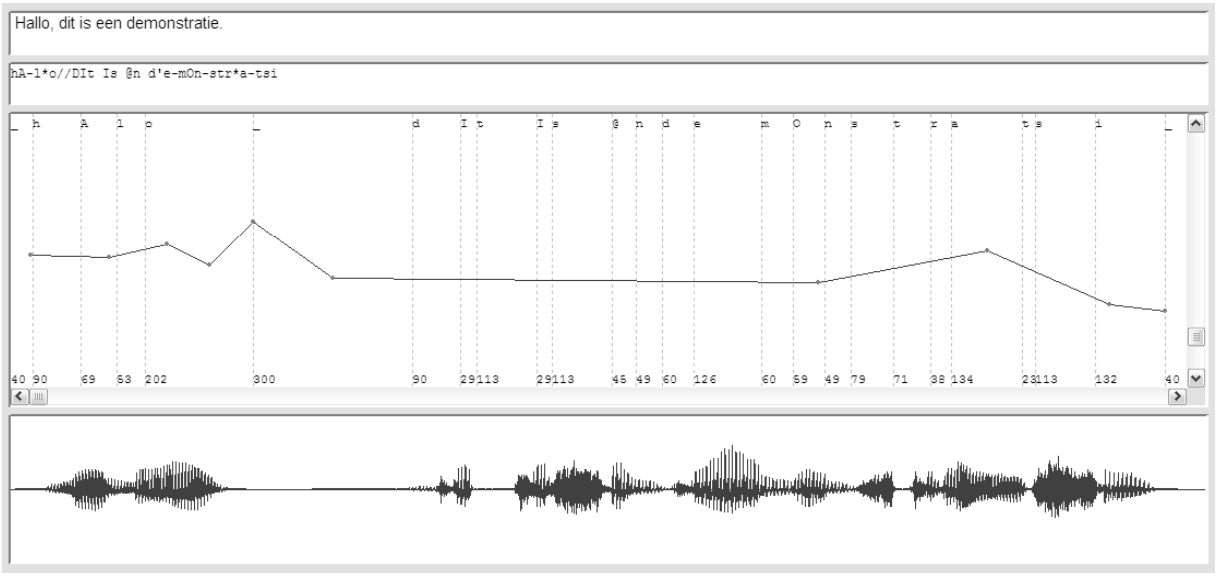

Figure 1. The Speech Editor application of Fluency TTS showing the intermediate stages in the generation of the sentence *Hallo, dit is een demonstratie* 'Hello, this is a demonstration'.

If the software is used to synthesize phonetic transcriptions of words in an electronic dictionary, the first stage is by-passed and the dictionary software provides the transcription. This is necessary, as the text-to-speech software cannot predict the pronunciation of ambiguous words such as *VOORkomen* 'happen' versus *voorKOmen* 'avoid' (small capitals indicate stress). It is also useful, as it allows us to define a more careful, emphatic pronunciation of the isolated dictionary words than is necessary for continuous speech. For example, the word *onAARdig* 'unkind' might be transcribed with a glottal stop before the second vowel in the dictionary, whereas for text-to-speech it might be preferable to have the /n/ in the first syllable resyllabify with the second syllable.

Regarding the third stage, a developer of text-to-speech software has a choice to either generate the audio from scratch (using articulatory or formant synthesis), or concatenate recorded speech samples (usually diphones). Excellent results can be obtained both ways, but with the latter approach it is easier to achieve the degree of naturalness required for commercial application.[2] The Fluency software uses MBROLA diphone concatenation, with diphones for both a male and a female voice, and approximately 2500 diphones for each voice. Included are diphones for French nasalized vowels (e.g. *ch<u>an</u><u>on</u>*), so the many French loans in the Dutch vocabulary can be synthesized correctly.

4 Phonological rules

In the second stage of the conversion of text to speech in the Fluency software, prosody rules are applied to the phonetic transcription to derive a specification for the synthesizer. In our approach, the term prosody applies not only to durations and intonation, but also covers

[2] With formant synthesis one can control almost every aspect of the synthesis process. This makes it rather suitable for phonetic experimentation (e.g. Dirksen & Coleman, 1997).

allophonic variation and phonological rules, which are integrated with the duration rules. The rules do not implement a full phonology of the Dutch language[3]: some of the effects of phonological rules are implicitly present in the diphones, other effects are assumed to be present in the phonetic transcriptions. In each case, we have looked at what was needed (for example, rules which operate across word boundaries need to be dealt with), and how we could obtain the best result in terms of speech output.

For example, there is no need to implement Homorganic Glide Insertion for words such as *du[w]o* 'duo' and *be[j]o* 'beo', or to require that these glides are included in the phonetic transcriptions. The reason is that these intervocalic glides are implicitly (and systematically) present in our diphones. On the other hand, we did need to implement Schwa Epenthesis for cases such as *mel[ə]k* 'milk' and *ker[ə]k* 'church', as the diphones for our male voice are somewhat infelicitous in this respect. However, we only needed a partial implementation of the rule, as many cases, for example *her[ə]fst* 'autumn', sound excellent without an epenthetic schwa[4]. In cases where the insertion does take place, we make the schwa very short, to account for the fact that it is part of a particular transition rather than a full segment.

We also found no necessitiy for Nasal Assimilation. Within morphemes, as in *ba<u>n</u>k* 'bank', we assume that a velar nasal is indicated in the phonetic transcription. Across morphemes, e.g. *i<u>n</u>komen* 'income' or *be<u>n</u> klaar* 'am ready', the diphone for [n-k] supplies a sufficient amount of natural assimilation (see also Jongenburger & Van Heuven, 1993). In fact, such natural assimilation is to be preferred in our view, as it provides a more subtle hint of what is going on than can be expressed in terms of phoneme symbols or phonological features.

In fact, when we cut our diphones, we tried to include these natural assimilations as much as possible. As a general rule, coarticulation between two consecutive sounds A and B is anticipatory in nature, that is, A will adapt to B much more than B to A. This can easily be seen in the spectrogram for a vowel which is in between consonants. The consonant-to-vowel transition shows very rapid formant transitions. In the vowel-to-consonant transition, on the other hand, formant transitions are slow, and they start early in the vowel. As a result, very early on in the vowel (as soon as the vowel targets are reached), very little evidence remains of the preceding consonant, but the effects of the following consonant can be seen (and heard) early on. So, although the common rule of thumb for cutting diphones is to cut somewhere in the middle of a sound, we prefer to cut very early in most sounds, so coarticulatory effects are taken into account as much as possible. Plosives are a notable exception, in that here we cut directly before the burst, i.e. very late rather than early in the time course of the segment, because the silent interval, whether voiced or not, is coarticulated with the preceding sound. As a result, in the sequence [V-n-k] the vowel-to-nasal diphone will signal nasality, and the nasal-to-plosive diphone will signal velarity, which is just right for these assimilations.

For the same reason, Regressive Voice Assimilation, which says that obstruents are voiced before a voiced plosive, can safely be left in the hands of the diphones. If the word *ee<u>t</u>baar* 'edible' is synthesized with a /t/ in the transcription, the result is more subtle than when a /d/ is transcribed, as the latter sounds overly casual (especially for a dictionary). A duration rule for intervocalic consonant clusters will make the [t] rather short, which also helps to indicate

[3] We have used Booij (1995) as our main reference for Dutch phonology.

[4] The diphone transition from a (rolling) [r] to [f] provides a subtle hint of a schwa.

an assimilated voiced plosive, as voiced obstruents are much shorter in duration than their voiceless counterparts. Finally, it is interesting to note that in clusters of two plosives there will generally be one (assimilated) burst rather than two identifiable segments.

However, perhaps as a result of our strategy for cutting diphones, we do need to implement Progressive Voice Assimilation, which says that a fricative is devoiced after a voiceless obstruent, both within and across words. And we do need to take care of interactions between the voice assimilation rules and Degemination.

The Degemination rule says that two consecutive identical consonants merge into one. This happens at morpheme boundaries, e.g. *nachttrein* 'night train', but also across words, e.g. *hij kan niks* 'he can do nothing'. Rather than merely deleting one of the two, we make the remaining one almost twice as long, so it still signals it twin brother. Again, this sounds more subtle and careful than a single consonant, and it provides an interesting example of the interplay between phonology and phonetics.

In derivational phonology, the output of the voice assimilation rules may feed Degemination, as in *afval* 'garbage' and *klapband* 'flat tire'. Although the [f-v] and [p-b] diphones do not sound too bad in these cases, we did incorporate these assimilations in our rules. However, as our rule system is not derivational, we had to adapt our Degemination rule with special cases for these assimilated obstruent clusters.

Our rules are also sensitive to the extra information encoded in the Van Dale transcriptions (section 2). The secondary stresses and compound boundaries are a useful guide to duration rules that define the rhythm of a word or sentence. Also, the special nature of the plural suffix /n/ in *komen* 'come' is dealt with by our rules. If it resyllabifies, as in *ze komen en gaan* 'they come and go', it is treated as any other onset /n/. If it does not, it is not deleted (as a straightforward phonological rule would demand), but given very short duration (less than 15 ms.). Such a short [n] is not perceived as such, but presents itself as a hint of nasality on the preceding schwa, which sounds more refined than a simple deletion, or, for that matter, a full [n], which sounds overarticulated or even wrong.

The fact that we do not always need to explicitly define a phonological rule in our software, does not mean that we cannot do so, should this be needed. All rules in our system can be parameterized, and these parameters can be changed at run-time. In an early version of the rule set, the explicit application of phonological rules could be selected by the end-user. The idea was that explicit assimilations and deletions define a somewhat casual speaker, whereas implicit, natural assimilations, and reductions instead of full deletions, define a more formal, refined speaker, which the end user might expect from a talking dictionary. This approach allowed us, for example, to synthesize three versions of the word *postbode* 'mail man' (see section 2):

- a version with a rather short [t] (about 40 ms, a single intervocalic [t] receives a duration of 90 ms).
- a version without the /t/, but with natural assimilation of [s] to [b].
- a version without /t/, and with explicit assimilation.

We found that we could amuse an audience of linguists and phoneticians with such demonstrations, but for most end users the differences were just a bit too subtle. So, in the end, we decided to remove the clutter of if-then-else's and select the more refined speaking style as the only option.

However, one optional rule remained, as users seemed to have rather pronounced preferences. It concerns the allophony of /r/, which can be a rolling [r] or a more vowel-like [R], which is not unlike /r/ in American English. In our analysis of Dutch, the [r] is selected if it appears directly before a vowel (and in cases of Epenthetic Schwa like *ker[əɹ]k* 'church' or *her[əɹ]fst* 'autumn'), whereas the [R] is typical of the coda position. But many users prefer a rolling [r] in all positions, so we offer this as an option.

5 Sentences

It is one thing to hear a word spoken in isolation, quite another to hear the pronunciation of that word in a sentence. A dictionary usually gives one or more examples of the usage of a word, and of how the word may be used in idiomatic expressions. As an extra service to the user, an electronic dictionary may offer audible pronunciations of these sentences as well. With the use of text-to-speech software, this can easily be achieved. It is even possible, and especially useful for non-native users, to highlight each word as it is being spoken (a standard feature of modern text-to-speech software). In this case, the input to the speech synthesis software is not a phonetic transcription but plain text.

However, if the text-to-speech software has to decide on the pronunciation of a sentence all by itself, it cannot be guaranteed that it will always be correct. Ambiguous words present problems that cannot always be solved by the text-to-speech software, even if it uses contextual analysis. Also, the distribution of pitch accents in a sentence is a thorny problem which involves not only syntactic structure, but also the (presumed) discourse context (Quené & Kager, 1993; Dirksen & Quené, 1993; Hoekstra, 2004).

The most straightforward solution to these problems is to (have the editors of the dictionary) annotate sentences where necessary. Ambiguous words can be solved by giving the exceptions a special code and providing additional lexicon entries with the correct pronunciation of these forms. Some examples:

hij kon niet voorkomen dat het fout ging 'he could not prevent it to go wrong'
het kan voorkomen\1 dat de deur gesloten is 'it can happen that the door is closed'

het regent de hele dag 'it rains the whole day'
de regent\1 is afwezig 'the governor is absent'

In the first example, the code '\1' is used to refer to an extra lexicon entry with the pronunciation for the word *VOORkomen* 'happen', with stress on the first syllable rather than the second (the speech synthesis software selects *voorKOmen* 'prevent' as the default). In the second example, the software knows the more frequent *regent* /rˈe-ɣənt/ 'rains', but not the rather arcane *regent* /rə-ɣˈɛnt/ 'governor', so the latter is added to the lexicon with a special code.

Pitch accent patterns can be indicated to the Fluency text-to-speech software by using the following tags: '\+' to add a pitch accent to a word, '\–' to remove a pitch accent, and '\<' to indicate a rhythmic stress reversal. Some examples (accented syllables are capitalized):

de tele**VI**sie is de **HE**le \–avond niet \+**AAN** geweest 'the TV has not been on all evening'
ik heb het \<**HE**lemaal met hem \+ge**HAD** 'I've really had it with him'

In the first example, *aan* 'on', which is in the synthesizer's list of unstressed function words, is used as a particle and should receive the main focus of the sentence. A pitch accent on *avond* 'evening' is not wrong, but the sentence is rhythmically much better without it. In the second example, the normal stress pattern of *helemaal* 'really' is with the main stress on the final syllable, and a secondary stress on the first. The '\<' tag reverses this pattern to improve the rhythm of the sentence.

Using these tags, a collection of examples and idiomatic expressions can easily be optimized so they are spoken with adequate prosody by our text-to-speech software. Example sentences are usually short and simple, so most will not need any modification at all. The extra codes can easily be hidden by the dictionary software.

6 Conclusion

Text-to-speech software can be used in electronic dictionaries to demonstrate to the user the pronunciation of words and example sentences. We have seen that some amount of experimentation and fine-tuning is both useful and necessary. Phonetic transcriptions embedded in the dictionary software should not attempt to provide pronunciation details that can be added with more subtlety by the speech synthesis software (section 4). On the other hand, the text-to-speech software sometimes needs a helping hand with the pronunciation of sentences (section 5).

The close collaboration between Fluency and Van Dale Lexicografie has resulted in high-quality text-to-speech software, that is used as a stand-alone product, but also as the pronunciation module in several of Van Dale's electronic dictionaries.

References

Booij, G. (1995). *The Phonology of Dutch*. Oxford: Oxford University Press.

Dirksen, A., & Quené, H. (1993). Prosodic Analysis: The Next Generation. In V. J. van Heuven, & L. C. W. Pols (Eds.), *Analysis and Synthesis of Speech* (pp. 131-144). Berlin: Mouton de Gruyter.

Dirksen, A., & Coleman, J. S. (1997). All-Prosodic Speech Synthesis. In J. P. H. van Santen, R. W. Sproat, J. P. Olive & J. Hirschberg (Eds.), *Progress in Speech Synthesis* (pp. 91-108). New York: Springer.

Dutoit, T. (1997). *An Introduction to Text-To-Speech Synthesis*. Dordrecht: Kluwer.

Froon, J., Den Hartog, J., & Zuidema, J. (1999). Beter verbeteren: slimme spellingchecker. *Natuur en Techniek*, *67* (12), 7-15.

Jongenburger, W., & Van Heuven, V. J. (1993). Sandhi Processes in Natural and Synthetic Speech. In V. J. van Heuven, & L. C. W. Pols (Eds.), *Analysis and Synthesis of Speech* (pp. 261-276). Berlin: Mouton de Gruyter.

Hoekstra, H. (2004). (De-)accenting and discourse structure. In this volume.

Quené, H., & Kager, R. (1993). Prosodic Sentence Analysis without Parsing. In V. J. van Heuven, & L. C. W. Pols (Eds.), *Analysis and Synthesis of Speech* (pp. 115-130). Berlin: Mouton de Gruyter.