Introductory Evaluation of the Swedish RealSpeak System

Jonas Lindh
GSLT (Graduate School of Langauage Technology) 2001
Term Paper for Speech Technology 1

Abstract

The development of different text-to-speech systems have evolved massively the last decade and new techniques have been introduced that have taken the technology further.
Lenout& Hauspies RealSpeak is a diphone-concatenation system with automatic unit selection. What differs it most from other systems is that it uses raw speech as output. It also does not only have one instance of each diphone, it uses quite a large speech corpus, which contains hundreds of instances.
Given a phoneme stream and a target prosody for an utterance, it selects an optimum set of acoustic units which best match the target specification.
After working with the Swedish system there are a lot of subjective conclusions made, but an investigation about the bugs of the system and an analysis can help us to understand some of the problems with a natural sounding synthesis and maybe give some clues on how to deal with them.
This article will deal with:
- How does the Swedish Realspeak system perform?
- What parts does it not handle well and why?
- Is this technology something to build on for future text-to-speech systems and how in that case?
An evaluation is needed since the technology used is far from perfect, but maybe the best so far. This also means that building new systems from this technological platform might give us an even more natural sounding synthesis, which is needed for a lot of applications.
Tests will be based on subjective ratings from the demo at www.lhsl.com/realspeak and spectrogram analysis using Wavesurfer and Praat. Spectrograms can be found in the appendix.

Background

An introduction to unit selection systems is given in the article *Speech Processing for communications: What´s new?*

> " The acoustic difference between units chosen for concatenation and these units as they would appear if the sentence had been read by a human. The cost is termed *target cost*, for obvious reasons. It accounts for differences in timber, duration, and intonation. Of course the synthesizer does not know how units would sound if the sentence was pronounced by a human reader (if it would, the synthesis problem would be solved), so this cost is only *estimated* (most often with phonetic-based predictors).
> - The continuity between concatenated units. It is always better to have units that lead to smooth speech when concatenated. The related cost is termed as *function cost*. It accounts for differences in spectral envelope and intonation. This cost can be estimated from an acoustic analysis of units, using more or less sophisticated distance measures (the more perceptual the distance measure, the better)."
> (S. Deketelaere, O. Deroo, T. Dutoit, 2000)

This gives an overview on how systems like these work. There are still just a few systems like this, probably because it is expensive to produce them and also time demanding. Recordings of decent quality must be done

and these must be segmented and labeled. Lexicons must be built with transcriptions and a letter to sound rule system has to be implemented to take care of words that do not appear in the lexicons. An engine that selects the best suited diphones from the recorded material must be built and a decompounder should be built to be able to label the correct stress to unknown compound words. Methods for autmatic labeling and segmentation are getting better though, and in the future this might mean less manual work. When the system is built, new languages can be implemented in the same by just exchanging the different parts suitable for the langauge. Another idea is to implement new voices which might seem simple by just exchanging the diphone database. The problem is that a system like this is very speaker dependant which will be discussed further below.

1. Parts of a Natural Sounding System

In this chapter there will be a thorough explanation about the different parts of the Swedish RealSpeak system and their functions.

1.1 Diphone Database

Existing diphones of raw speech were recorded in a studio with a female speaker born and raised in Stockholm, Sweden.

> "… diphones are often used in speech synthesis as they provide a reasonable balance between context-dependency and size (typically 1000-2000 in a langauge). In speech synthesis, diphone units normally start half-way through the first phone and end half-way through the second. This is because it is known that phones are more stable in the middle than at the edges, so that the middles of most /a/ phones in a diphone are reasonably similar, even if the acoustic patterns start to differ substantially after that. If diphones are concatenated in the middles of phones, the discontinuities between adjacent units are often negligible."
> (D. Jurafsky and J. H. Martin, 2000)

This is also the way it is done for RealSpeak, even though stops are divided into closing phase and burst.
As shown by Öhman (1966) there can also be assimilation between vowels over a consonant which could create problems in concatenating diphones, but by using as large units as possible the concatenation can be successful.

Vowels and diphthongs

To make a cut in the middle of vowels can be constrained in the engine (see below). The problem is that it might also makes the system constrained to use a certain set of diphones. Constrained choices will make the system perform worse in handling sound combinations such as uncommon diphthongs.

> "The acoustic theory of vowel production (Chiba and Kajiyama, 1941; Fant 1960; Stevens and House, 1961) showed that vowels can be represented by an all-pole vocal tract transfer function, and that the relative amplitudes of the formant peaks can be predicted from a knowledge of formant frequencies, as long as the vowel is not nasalized."
> (Dennis H. Klatt, 1986)

Nasalized vowels often appear after, or in between nasals. This makes the formants look kind of blurry, but in listening tests it was obvious that this almost never created problems since the nasalized vowel never was selected and connected to one that was not. Lip protrusion lowers all formant frequencies (most in F3) in neighboring sounds. This problem was consistent, and in all test sentences there are constantly frequency jumps

in rounded sounds such as /y/ and /ω/. Bad concatenation is usually avoided though through measurements over the three lowest formants before selecting and connecting units.
It is difficult to choose which diphthongs a Swedish system should handle considering foreign names and other common (or less common) pronunciations. “New York” and “Houston” in Swedish for example. See *Spectrogram 1.1* of the sentence “*New York är en trevlig stad, men Houston är vackrare”*. While analysing recording it was also discovered that there were almost no instances for diphthongs. Almost all chosen units for diphthongs was taken over word boundaries.

How to handle phonemes like /h/ and stops

> "The consonant /h/ is sometimes grouped with the fricatives because it is noise-excited, but /h/ functions more like a voiceless sonorant consonant. The sound source for /h/ is aspiration generated near the larynx, the vocal tract assumes the shape of the following vowel, and all formants are weakly excited by the noise."
> (Dennis H. Klatt, *Review of text-to-speech conversion for English*, 1986)

Even though the phoneme /h/ is considered a voiceless consonant in isolation, it seldom appears that way in context. Cutting the phoneme in half can make a lot of formant transitions uneven. In *spectrogram 1.2* of the sentence “*Har hästar hunnit häva havre hungrigt, haha”*, it is possible to see how some of the /h/ phonemes that are cut in half have strange transitions to the following vowel sound and when listening to it, it is easy to hear where the transitions are not smooth enough to create a natural sounding speech. Restraining the system not to make cuts within certain phonemes (like /h/) can restrain the system, but the only way to handle a phoneme like /h/ is probably to avoid cutting as much as possible by recording it in enough contexts and restrain the engine from splitting it.
By splitting stops in closing phase and burst a lot of problems are avoided. The burst diphone contains the first transition to the following sound, if it is cut to early in the following segment and connected to a vowel from a different context the transition is not always smooth. See *Spectrogram 1.5* of the sentence “Bor biet Bob i bur, by, eller bör bi bära borr och be i bar”.
Unvoiced stops are usually followed by aspiration, unless it is preceded by a sibilant. Some burst were taken from this context and concatenated with a stop that was supposed to have aspiration. This was only taken care of in the engine, but there are still cases where the aspiration is lost.

Stress levels labeled

There are three levels of stress labeled, primary stressed syllables /’bo:rd/, secondary stressed syllables /kon’si:s/ and unstressed /at/. How a word is stressed in a sentence is chosen by the person segmenting the speaker input. Since the grapheme to phoneme conversion (G2P) basically put stress on all content words it would be strange to label them differently, since they will not be used in the concatenation procedure otherwise. To get a smooth concatenation of the diphones in a speaker dependant system like this it is crucial to constantly update the lexicons and G2P (see below) not only according to common standard Swedish phonology, but also depending on the way the segments are labeled, which hopefully is done in a consequent way. In early versions of the system (for example the one that the Linguistic Department in Gothenburg uses) there are mistakes like putting the secondary stress on the second syllable in the word “talsyntes” (Speech Synthesis), which is a word most likely to be used often. It was discussed if more stress levels could be labeled to restrain the selection, but tests showed that instead of getting less cuts there was bad concatenations and more fundamental frequency “jumps”.

Labeling and segmentation of recorded sentences

The segmentation procedure was done manually using spectrogram and a labeling sheet. Coarticulation of some sounds make it impossible to cut the specific phoneme in half (as mentioned above).

To be able to label segments properly and then use them efficiently, some allophones and optional features were added to the original phoneme table. The allophone schwa was added for example to improve pronounciation of unstressed /e/ and /ɛ/. Glottal stops were intoduced as an additional feature since the speaker used them to emphasize a stressed syllable if the preceeding and following sound was a vowel. Creaky voiced segments were deleted from the speechbase unless it was the last segment of an utterance.
By using the inherent qulities of recorded raw speech one can extract the most important parameters like formant frequencies and concatenate the best suited units.

1.2 Lexicons and transcriptions

RealSpeak contains different lexicons put together with transcriptions of Standard Swedish pronunciation.
All these lexicons are converted from different phonetic representations to a TTS transcription made up by the company. As most of us know the IPA transcription is not perfect to describe all languages and neither is this.

Phonetic Transcription of Swedish

There was at first no way to describe the Swedish accents (grave and acute) and since the accents differ words the pronounciation of the words “pålen” and “Polen”, which is a minimal pair, is important for understanding utterances. So a way to label them was made up. When all syllables were labeled with accent 1 and accent 2 there was a lot less diphones to pick from in the speech base. The second choice would be to double the speechbase with all kinds of occurences of accent 1 and accent 2, which ofcourse would be too space demanding. What could be done was to make the system prefer one accent over the other. If it is given a lower weight node in the engine (see below) to connect the same accent in a word it might make the selection poorer, but at least in most cases a misleading pronunciation could be avoided.

Bugs

A problem with different lexicons put together with several different people working on it is that they have all different references to the transcription. A lot of mistakes or inconsequence in transcriptions lead to high mismatching when selecting units which were labeled differently from the way they were transcribed in the lexicons or by the G2P (see below). For example the system had problems choosing between the short /e/ and /ɛ/ phoneme. Not only because of the inherent qualities of Swedish phonology, but because it was transcribed and labeled in different ways.
In *spectrogram 1.3* of the sentence “Säger eva e, i ett, eller ä, i ätt”, there is no quality difference between the phonemes /e/ and /ɛ/. There is not much difference between the two phonemes either, but if they are labeled some times as /e/ and sometimes as /ɛ/, there will not only be a misleading pronounciation sometimes, but also less units to choose from.
It might be very interesting to have a lot of research done on which transcription to use, but for the system itself it was shown that if the lexicon transcriptions, G2P rules (see below) and labeling was done in a consequent way the system performed its best. (Authors remark from report written at L&H)
How should a system handle transcription of diphones that cannot be found in the database? As shown in *spectrogram 1.1*, the diphthong /ou/ in “Houston” is pronounced with silence. Another way to handle this is to make the system choose a neighboring sound and use that as a substitute or simply exclude one sound and just have a smooth transition, which a human reader with no knowledge of the pronounciation probably would do.

1.3 G2P (Grapheme to Phoneme conversion)

Handles all the cases when words cannot be found in the lexicons.
Trying to generate written strings into a phonetic transcription is challenging if all of Swedish phonology is supposed to be covered. The G2P was built by different modules. It contains a decompounder for compound

words, rules for assmililation and reduction over word boundaries, standard conversion for graphemes to phonemes in different contexts etc.
The G2P work was done from a set of rules taken from all kinds of information about Swedish phonology that could be found. But in a unit selection system like Realspeak all rules are speaker dependant. This means that if a common phonological rule is generated, the speaker might act in a completely different way. If there is a thorough phonological research done on the recordings, there can be a set of rules written down that later can be implemented according to the unique phonology of the speaker. This would have made the system use larger clusters of speech that was recorded. A comparison of what words the G2P can handle in the lexicons is needed to be able to decrease the lexicons, avoid overgeneration and a more space demanding system then what is actually needed.

Assimilation and reduction

To be able to concatenate the best way you need to handle assimilation and reduction not only in a phonetic knowledgeable way, but a speaker dependant way. If the speaker uses a certain way to pronounce an utterance in a recording, you really want to use exactly that part to be able to have the best concatenation of diphones.
Depending on the purpose of the system, which in this case is a natural sounding one, one can question whether a system is supposed to use a reduced pronounciation. In this case we got the best results while using the same kind of reduction and assimilations as the speaker did. For example "jag" is pronounced /ja:/, "och" is pronounced /o/ and "är" is pronounced /e:/.

Decompounder

Basic morphological rules were used to decompose compounds and move stress symbols to the right position. For a language like Swedish, which uses a lot of compound words, a tool like this increased the word intelligibility a lot when it was implemented. To make a system handle difficulties like compound words this needs to be built out though, because it still performs poorly in many cases. See *Spectrogram 1.4* of the sentence “Sammansatta ord är svårkontrollerbara”. The best way to save space would be if the decompounder could be built out, so that the huge lexicons can be decreased and the system easier to implement in applications.

1.4 Engine

The engine is used to pick the right diphone (or longer segments) from the database. Using a lot of different parameters makes it possible to get good concatenations even though it is space demanding. To get a closer look at what the engine contains the most important parts are described below.

PHONEME_ALPHABET

A description of the phonetic representations over all phonemes used for the system by one symbol per phoneme.

PHONEME_FEATURES_FILE

Describes :
field 1: phoneme identifier (1 char), gives the specific phoneme symbol that the system is identifying.
field 2: voicing type (voiced, unvoiced, medium), specifies voicing for the specific phoneme.
field 3: consonant/vowel distinction (C,V), tells the system whther it is a consonant or a vowel.
field 4: safe/unsafe cutting distinction (S,U), if a phoneme is Unsafe to cut it will get a lower value so that cutting in the specific phoneme is avoided (like /h/).
field 5: plosive/non-plosive distinction (P,X), specifies to the system whether a phoneme is a stop (plosive) or not.

field 6: phoneme groupings(N,V,F etc), specifies the group of phoneme to where the phoneme belongs like fricative, nasal etc. This feature is also connected to duration measurements.

OPTIONAL_FEATURES

Contains all optional features for compilation of a speech base, describing tones for Swedish grave and accute accent for example.

DATABASE_FILES

Gives the engine the database files compiled.

NODE_WEIGHTS

Gives features more or less prominence in the unit selection from the speechbase. For example syllable positition in a word left or right, syllable position in a sentence, word position in a sentence, final syllable position in phrase and pitch anchor will measure the fundamental frequency to avoid unnatural “jumps”.
There is no measurements done for amplitude. After some measuremants and comparison during the work with the system it was found that some of the worst clips was depending on difference in amplitude between two units put together. No tests were made with a node weight for amplitude continuity, which might have been a parameter to consider for further development.

TRANS_WEIGHT

Gives weights for features like phonemes adjacent in database, pitch difference in semitones at joint, vowel pitch and duration. There is also some measurements on tone continuity and cepstral distance to get the best suited diphones put together.

PHONE_VALUES

A value is given each phoneme to another, which it can be exchanged with if it can not be found in a certain position.

2. Word accuracy and other subjective ratings

To get some kind of idea on how the system performs, words were chosen randomly to see if the system pronounced them correctly.

Table 2.1 Realspeak Word accuracy in percentage of correctness.

Word type	Correct %	Comments	Incorrect %
1-syllable	100 %	"tös" no aspiration on /t/ Vowel discontinuities in "yr" and "uv"	0 %
2-syllable	100 %	Vowel discontinuities in "ljuga", "nysa" and "tutar"	0 %
Comp. Words	83,4 %	Incorrect words: "dödskul", "istapp", "nästippen", "polcirkel" and "rampfeber"	16,6 %

Handling one- and two-syllable words is (most of the time) not a very difficult task for RealSpeak, even though there are some discontinuities in the concatenation. Homographs like “banan” is not taken care of and is only transcribed as the word that occurs most often after statistical measurements.

3. Conclusion

As mentioned, the system does not handle compound words very well, but the biggest problem is how to concatenate units even better without having to record even more instances and make the system even bigger. To solve those problems there has to be better research on the recorded material, better labeling and lexicons. The G2P has to handle more then it does today so that one can get rid of the huge lexicons. There has to be more rules implemented to decrease the amount of discontinuities in concatenation, but also more units to choose from. This can be done by thoroughly investigating what units that needs to be recorded and make plans on how to extract them from read speech.

By using the inherent quality of recorded speech from one single speaker, the basic technology used in the development of RealSpeak performs very well. By building out grapheme to phoneme rules and implement effective and well working decompounders it is possible to build even better systems. Natuaral sounding synthesis can be used for many purposes. A future idea is to be able to exchange the diphone database quickly and have voice donors for speech disabled people. As most systems today are multilingual, it should also be possible to create new databases with different voice qualities. Maybe one for angry, happy or sad. This can be done with systems that can perform a lot of measurements automatically.

References

Deketelaere S., Deroo O. and Dutoit T. *Speech Processing for communications: What´s new?* at MULTITEL ASBL, Faculté Polytechnique de Mons 28, 91–107, 2000

Jurafsky D. and Martin J. H., *Speech and Language Processing*, University of Chicago, Boulder, 2000

Klatt Dennis H. *Review of Text-to-Speech Conversion for English*, Journal of the Acoustical Society of America Vol.82 s 737-793, 1987

Öhman S. *Coarticulation in VCV utterances: Spectrographic measurements, Journal of the Acoustic Society of America, 39* , 1966

Litteraure

Dutoit, T. *An introduction to text-to-speech synthesis* Dordrecht: Kluwer Academic, 1997

Carlson R., Granström B. *Speech Synthesis*, Hardcastle & Laver (editors) *The Handbook of Phonetic Sciences*, Blackwell Publishers Ltd, Oxford 1997, 768-788

Test Words

The test words were picked out randomly from a lexicon with help of a prolog program. The only criteria was number of syllables and variation of the first letter (mostly consonants).

One-syllable words:

"bil", "dis", "dass", "ful", "fin", "gråt", "get", "hur", "hiss", "jul", "kul", "kysk", "lat", "lim", "mask", "mus", "nös", "Nice", "pisk", "pack", "Quist", "rör", "riv", "söt", "snygg", "tös", "titt", "uv", "yr", "zon".

Two-syllable words:

"anden", "bilen", "cykel", dricker", "fjantig", "ganska", "gillar", "hinner", "harkla", "igel", "juvel", "knarka", "kista", "lista", "ljuga", "lama", "maska", "minnas", "nysa", "Nisse", "Opel", "peka", "pinne", "röra", "rinna", "springer", "satsar", "tittar", "tutar", "zoner".

Compound words:

"asbra", "bäddmadrass", "celsiusskala", "dödskul", "enbärsbuske", "fallucka", "genomtrevlig", "hårdvara", "istapp", "jättehärligt", "kanonkula", "lastgammal", "mansgris", "nästippen", "ostgratäng", "psykologstuderande", "Quistberg", "rampfeber", "snusförnuftig", "talsyntes", "telefonbanken", "underrepresenterad", "vuxenstöd", "valfångst", "Wallenberg", "yrväder", "årsinkomst", "ätbar", "överdriven".