Unit Selection and Emotional Speech

Alan W Black

Language Technologies Institute, Carnegie Mellon University &
Cepstral, LLC
awb@cs.cmu.edu

Abstract

Unit Selection Synthesis, where appropriate units are selected from large databases of natural speech, has greatly improved the quality of speech synthesis. But the quality improvement has come at a cost. The quality of the synthesis relies on the fact that little or no signal processing is done on the selected units, thus the style of the recording is maintained in the quality of the synthesis. The synthesis style is implicitly the style of the database. If we want more general flexibility we have to record more data of the desired style. Which means that our already large unit databases must be made even larger.

This paper gives examples of how to produce varied style and emotion using existing unit selection synthesis techniques and also highlights the limitations of generating truly flexible synthetic voices.

1. Background

Unit selection speech synthesis systems, e.g. [1], have shown a significant improvement in output voice quality. Selecting appropriate sub-word units from large databases of natural speech has raised the level of speech synthesis to a quality, in its best case, equivalent to that of recorded speech. The quality is directly related to the implicit quality and style in the recorded databases, and at last the voice output sounds like the original speaker (though this has been said before).

Since the publication of a well defined selection algorithm for unit selection, [2], we have seen significant new work in acoustic measures, and in alternative algorithms for optimally finding the best set of units to join together (e.g. [3]). However in the search for better algorithms, we have also noted that better databases that cover the acoustic phonetic space of the language in question can also make significant contributions to the quality.

In [2], the notion of **target cost** for a candidate unit from a database with respect to the required unit is presented in the following formula.

$$C^t(t_i, u_i) = \sum_{j=1}^{p} w_j^t C_j^t(t_i, u_i)$$

That is the target cost is a weighted sum of differences of features between the desired target unit and particular candidate unit in the database.

In addition units selected must not only have a small target cost but also join well. *Join costs* may be defined between two units as

$$C^c(u_{i-1}, u_i) = \sum_{k=1}^{q} w_k^c C_k^c(u_{i-1}, u_i)$$

The optimal selection of units is the set that minimize

$$C(t_1^n, u_1^n) = \sum_{i=1}^{n} C^t(t_i, u_i) + \sum_{i=2}^{n} C^c(u_{i-1}, u_i) + C^c(S, u_1) + C^c(u_n, S)$$

Where S denotes silence and $C^c(S, u_1)$ and $C^c(u_n, S)$ address the conditions at the start and end of the utterance.

As highlighted in more detail in [4], the overall cost can be reduced in a number of ways, that do not just involved changing the acoustic measurements.

We can limit the set of utterances we wish to synthesize to those whose costs are low. [5] carries this to an extreme where the synthesizer defines a domain and will not synthesize outside that domain. However within domain, the quality can be very high, and for many applications this solution is ideal.

We can design the database itself to better cover the intended acoustic space, so that there are less possibilities for bad joins [6]. Appropriately designed databases are important for not just “domain” synthesizers but general synthesizers too, as they too, are designed to cover an intended (though larger) space.

Thus current unit selection can work well, when the desired utterances that must be synthesized are appropriate for the database they are to be selected from. It is notable that attaining variation outside that database is hard, and rarely even attempted as any form of signal processing to modify the spectral and prosodic quality of the speech, typically degrades the quality or at least makes it less natural.

2. Emotional Speech

Unit selection techniques will provide synthesizers with the quality of the database they are built from. Thus we can synthesized various emotions if we record database of the appropriate type.

However, before we give some examples of this direction, it is worth better defining what is meant by emotional speech, and more importantly how we might actually use such synthesizers in applications.

Traditionally emotional speech is split in four groups: neutral, happy, sad, and angry (hot and/or cold anger). Various studies show that listeners can fairly reliably distinguish between happy and sad, though may confuse these with hot anger and cold anger in ambiguous situations. Testing output quality is hard, studies usually use lexically neutral statements so just the spectral and prosodic properties vary, while in real life situations, lexical issues and context probably are a bigger clue to the emotional state of the speaker.

The following experiment highlights how lexical choice influences human perception of voice characteristics.

In developing a child voice synthesizer, we specifically required a gender neutral voice. Our recordings were based on an adult voice-over actress with experience in performing child voices. When we first tested recordings from her with a group of potential users we found most people identified the voice as an adult pretending to be a child. However we noted that the sentence contents, designed for phonetic and metrical coverage are not typical sentences that would be spoken by children. It is difficult to imagine situations where a child might say.

> A sense of psychological certainty is no proof in itself of epistimelogical validity.

Thus on later tests we synthesized child specific utterances to test the perceived view of the voice.

> Are we there yet?
> Please read me my a story.
> Can't I do it tomorrow?
> ...

We also synthesized girl specific sentences, and boy specific sentences

> Can I go to the Mall with Kimmy?
> I like to go shopping for new clothes.
> When I grow up I want to help animals.
> ... Last weekend my Dad took me to a ball game.
> I'm starving, is there anything to eat?
> My Mom says I'm not old enough to watch Wrestling. ...

We played these utterances to parents, not familiar with synthesis, and rather than ask them the gender of the speaker, asked them to give us a suitable name and suggest the age of the speaker. Overwhelmingly all listeners give boy names when listing to the "boy" sentences, and girl names for "girl" sentences. However in general the listeners did consider the boy younger than the girl.

These informal tests show that people's perception of voice type is subtle, and content can easily overwhelm prosodic and spectral qualities of voices.

In our experience in building speech synthesis systems, these standard definitions of emotion are actually rarely requested by users. Though much more subtle notions of emotion and style are needed.

3. Recording in style

When considering building a unit selection synthetic voice, knowing the most likely usage pattern can make it easier to define the most suitable style for building a voice.

To explicitly show how the same speaker may use different styles, and the listener may require the different, we constructed a voice designed to deliver the weather. This is very much a limited domain voice with an well defined explicit vocabulary and templates. We constructed 100 sentences that gave full coverage of temperature range, outlook, wind speed and direction etc. Then we recorded the same set of sentences in two distinct styles:

Genki : from the Japanese word for healthy, upbeat.

News : direct "no-nonsense".

A typical generated sentence would be of the form.

> At 7 P.M., the temperature is sixty-eight degrees Fahrenheit. The wind is from the north, at eight miles per hour. The barometric pressure is thirty inches, and steady.

The output quality of each of these synthesizers is by any standards excellent, but the styles are different. (http://cepstral.com/demos)

When playing these two synthesizers to people we get different reactions. Although we only have anecdotal results, people who actually want to know the weather prefer the News-type synthesizer while people who wish to be impressed by high-quality synthesis prefer the Genki-style voice. Neither synthesizer, can be criticized for being unnatural, but the difference in style in which the information is delivered makes a significant difference in the listeners views.

4. Emphasis

Even when considering something as basic as emphasis in speech synthesis we quickly discover that our control over stylistic aspects of speech to be very minimal. When humans speak they use a number of different variations to denote emphasis in speech. These include phrasing, duration, F0 excursions, and power. Different speakers may choose to render emphasis with different combinations and even individuals may change their strategies in different styles of speech.

In Festival, [7], emphasis is implemented by rather naive rules. In SABLE [8] marked up words, emphasis is realized by inserting short pauses before and after the emphasized word, extending the duration, and intensifying the F0. In simple cases this is adequate, but is very crude and it is easy to find cases where it sounds unnatural. However in almost all cases it is clear that the synthesizer is emphasizing that word, but potentially in a non-natural way, especially in poly-syllabic words and phrases.

In order to improve the quality of such a basic speech variation as emphasis we tried explicitly recording examples of naturally emphasized speech. As we wished to use these recordings in a standard unit selection synthesizer we had to ensure that there was sufficient phonetic, metrical and prosodic coverage within the data bases.

Thus we took a database originally designed with such coverage. We used the techniques described in [6], to select sentences that optimally provided the best coverage based on an explicit acoustic model of the voice talent's speech. This database consists of 548 sentences selected from out-of-copyright books ([9]).

Then to address the coverage for emphasis, we labelled every other word in each sentence as emphasized. The voice talent (AWB), then read the sentences with emphasis on each word as marked. This was actually harder than expected. It is not easy to read a sentence and put natural emphasis on arbitrary words. This fact is important in elicitation of varying styles for unit selection databases. It is hard for a voice talent, even a trained one, to consistently deliver a desired style. When the request is something as unnatural as common emphasis on multiple words in the same sentence, the result may not always seem natural.

Each of the words to be emphasized were marked with an underscore

> _Allow me _to interpret _this interesting _silence.
> _Tarzan and _Jane raised _their heads.

These were automatically labelled and a cluster based unit selection synthesizer was built [10]. In the default case units of the same phone type are clustered using a CART method that indexes the clusters by high level features such as phone context, metrical structure etc. In this case we tagged each phone with an emphasis feature. Thus phones from emphasized words can only be used in the synthesis of emphasized words, while phones in non-emphasized words can only be used in the synthesis of non-emphasized words.

Once build, we took a number of short sentences, not in the original database, and synthesized sentences emphasizing each word in turn.

_This is a short example.
This _is a short example.
This is _a short example.
This is a _short example.
This is a short _example.

In all cases it was easy to identify the emphasized word in the synthesized phrase, however in about 15% of the examples the emphasis was judged to be unnatural. Though other problems with this fully automatically built unit selection system do partially interfere with this result.

However despite the limitations of this particular database it is clear that this technique does work. If you record appropriate data with sufficient coverage it is possible to synthesize that style in a natural way.

5. Style

In our work on providing speech synthesizer for applications we have found that the wider notions of emotion are rarely requested. However particular styles have often been required for the applications we have worked with.

In our work in providing voices for the AAC market (Augmentative and Alternative Communication) where people use hand held devices to speak having lost (or never had) the ability to speak for themselves, style is very important as the synthetic voice becomes the persons own voice. Synthesizers based on news reader style speech such as the Boston University Radio Corpus [11], produce voice output that still sounds like a news reader. An AAC device is primarily used for dialog, rather than extended monologues therefore we took this into account both in instruction to the voice talent while recording, and in the design of the utterances to record.

Delivery style is crucial in voice recording. In the recording of canned prompts, it is said that the most common phrase said by the voice coach is “Say it again with a smile.” Like the Genki vs News style weather described above style in delivery defines the style of the synthesizer. Putting people in a small recording studio for hours on end and getting them to read thousands of sentences may be one reason why synthesizers often sound bored.

In the recent DARPA-funded Babylon project where we were part of a team to developed a two-way speech-to-speech translation system running on a standard PDA. Our Speechalator system offers English to Arabic and Arabic-to-English in the medical interview domain.

Apart from the non-trivial problems of running on such a limited platform, such systems require the voice output style to be appropriate for the message being delivered.

The first issue in style in speech-to-speech translation is that some utterances are commands, such as “Put down your weapons” while others should be delivered in a more compassionate style, such as “Where does it hurt?”. Inappropriate style for either of these utterances will be detrimental to communication. On an earlier speech-to-speech system developed by us [12], we did not take such care and the delivery of commands in the Croatian synthesizer were consider somewhat amusing by native speakers rather than as actual commands.

6. Recording in multiple styles

[13] identify two basic methods for dealing multiple styles in a unit selection speech synthesis paradigm. Separate voices can be built for different styles or domains, such as a command voice and an interview voice, and these voices may be switched between by the application using the synthesizer. This is called *tiering*. This technique works well when there is a well defined distinction between the voice types. For example, when the domain changes in a well defined way, weather information to flight information, or even good weather to bad weather information.

The second method for combining voice types is called *blending*. In this model the databases are mixed into the same database. This allows a more gradual changed between voice types, and the potential of mixed styles. The style selection is automatic based the requested units. This may be influenced implicitly by the words and phrases being synthesized, command words would be more likely to be synthesized from the command phrases in the database, while general information may come from a more neutral part of the database.

This technique works well in mixed domain based synthesis with other domain based databases and/or general ones, though it helps if they are basically in the same style. Mixing domain-based and general databases in a blended voice can produce excellent quality when in domain and reasonable quality when not, which is useful for many applications.

7. Recording all styles

There have been attempts to record very large databases of natural speech: either large amounts of data in the same basic style [14] or large mounts of data collected in different situations thus in varied style [15]. Such recordings are non-trivial tasks and a substantial amount of acoustic normalization is required in order to allow them to be used reasonably for sub-word unit selection. If the databases are not appropriately normalized joins will be very obvious when units are selected from different parts of the database or selection will be limited by the different recording conditions, style of speaker etc.

Recording all styles naturally would take a very long time, while prescribing styles is also very hard. When one voice talent delivered a 120 utterance shouting database, it was hard for them to speak normally for the following two days.

It is clear that current unit selection techniques work very well for limited styles and for particular applications this may be sufficient, but it is clear unit selection in its current state does not give us the flexibility we have in a human voice.

8. Conclusions

In order to get both the flexibility and naturalness of human speech in a synthesizer, it is clear we need to look closer at how we build our voices. Recording everything is not sufficient and already we are finding that recording very large databases is significantly hard.

When coverage problems like this exist in other fields the solution is to decompose the system so each part can be covered separately. For example we could consider separate spectral models, intonation models and duration models. This is some sense what was done with earlier diphone systems, and we know that these do not have the naturalness of unit selection systems.

We have however seen unit selection techniques being used on separate streams of information. For example [16] move towards unit selection techniques for selecting appropriate ToBI ([17]) labels for databases of intonationally labelled speech. [18] select F0 contours for databases of speech in the same basic way as (spectral) unit selection.

There is a disadvantage though, by decomposing the signal, we introduce the problem that we have to reconstruct it afterwards. The artifacts that such reconstruction introduces were one the reasons unit selection with minimal smoothing became popular.

But now that we are finding the limitations of conventional unit selection techniques, improving the decomposition and reconstruction of the signal, which would allow us to model component separately, seems like the most direct way to improve the flexibility of synthetic voices.

9. Acknowledgements

This work was partially funded by NSF grant "ITR: Prosody Generation for Child Oriented Speech Synthesis". The opinions expressed in this paper do not necessarily reflect those of NSF. The author would also like to acknowledge help from Kevin Lenzo of Cepstral LLC in discussions related to this work.

10. References

[1] M. Beutnagel, A. Conkie, J. Schroeter, Y. Stylianou, and A. Syrdal, "The AT&T Next-Gen TTS system," in *Joint Meeting of ASA, EAA, and DAGA*, Berlin, Germany, 1999, pp. 18–24.

[2] A. Hunt and A. Black, "Unit selection in a concatenative speech synthesis system using a large speech database," in *ICASSP-96*, Atlanta, Georgia, 1996, vol. 1, pp. 373–376.

[3] E. Klabbers and R. Veldhuis, "On the reduction of concatenation artefacts in diphone synthesis," in *ICSLP98*, Sydney, Australia., 1998, pp. 1983–1986.

[4] A. Black, "Perfect synthesis for all of the people all of the time," in *IEEE 2002 Workshop on Speech Synthesis*, Santa Monica, CA., 2002.

[5] A. Black and K. Lenzo, "Limited domain synthesis," in *ICSLP2000*, Beijing, China., 2000, vol. II, pp. 411–414.

[6] A. Black and K. Lenzo, "Optimal data selection for unit selection synthesis," in *4th ESCA Workshop on Speech Synthesis*, Scotland., 2001.

[7] A. Black, P. Taylor, and R. Caley, "The Festival speech synthesis system," http://festvox.org/festival, 1998.

[8] R. Sproat, A. Hunt, M. Ostendorf, P. Taylor, A. Black, K. Lenzo, and M. Edgington, "SABLE: A standard for TTS markup," in *International Conference on Spoken Language Processing*, Sydney, Australia, 1998.

[9] M. Hart, "Project Gutenberg," http://promo.net/pg/, 2000.

[10] A. Black and P. Taylor, "Automatically clustering similar units for unit selection in speech synthesis," in *Eurospeech97*, Rhodes, Greece, 1997, vol. 2, pp. 601–604.

[11] M. Ostendorf, P. Price, and S. Shattuck-Hufnagel, "The Boston University Radio News Corpus," Tech. Rep. ECS-95-001, Electrical, Computer and Systems Engineering Department, Boston University, Boston, MA, 1995.

[12] A. Black, R. Brown, R. Frederking, R. Singh, J. Moody, and E. Steinbrecher, "Tongues: Rapid development of a speech-to-speech translation system," in *HLT2002*, San Diego, California, 2002, pp. 2051–2054.

[13] K. Lenzo and A. Black, "Customized synthesis: blending and tiering," in *AVIOS2002*, San Jose, CA., 2002.

[14] H. Kawai and M. Tsuzaki, "A study of time-dependent voice quality variation in a large-scale single speaker speech corpus used for speech synthesis.," in *IEEE 2002 Workshop on Speech Synthesis*, Santa Monica, CA., 2002.

[15] N. Campbell, "Towards a grammar of spoken language: Incorporating paralinguistic information," in *ICSLP2002*, Denver, CO., 2002.

[16] S. Pan, *Learning Intonation Rules for Concept-to-Speech Generation*, Ph.D. thesis, Columbia University, 1998.

[17] K. Silverman, M. Beckman, J. Pitrelli, M. Ostendorf, C. Wightman, P. Price, J. Pierrehumbert, and J. Hirschberg, "ToBI: a standard for labelling English prosody.," in *Proceedings of ICSLP92*, 1992, vol. 2, pp. 867–870.

[18] F. Malfrere, T. Dutoit, and P. Mertens, "Automatic prosody generation using suprasegmental unit selection.," in *Proc. ESCA Workshop on Speech Synthesis*, Australia., 1998, pp. 323–327.