Speech Synthesis for Educational Technology

Alan W Black

Language Technology Institute,
Carnegie Mellon University, Pittsburgh, PA
awb@cs.cmu.edu

Abstract

This paper gives an overview of the present state of the art in speech synthesis and its relationship to spoken output in education systems. The paper specifically looks at the use in general tutorial systems, use in language learning and supporting new languages, and in voice conversion techniques that can produce speech similar to a specific speaker.
Index Terms: speech synthesis, education

1. Introduction

Speech Synthesis technology has improved greatly over the last decade such that the quality of speech output in many applications can approach that of prerecorded speech. The development of **unit selection** speech synthesis [1], where appropriate sub-word units are selected from large databases of natural speech, has provided a comparatively easy method for generating high quality voices.

Previously the technology required carefully crafted database of natural speech, where specific phonetic coverage was explicitly designed into the databases [2]. This collectively is referred to as **diphone** technology, and is still used in some older systems. Although it has the understandability of later systems, it does not provide the naturalness of unit selection systems.

The naturalness of unit selection systems depends greatly on the databases recordings. Thus designing appropriate databases becomes critical to the quality of the output voice. But with this high quality comes a price. As the quality of the voice approaches human quality, a slight deviation in quality can be more unacceptable in a high quality voice than in a medium quality voice. Also because there are many possible concatenation points in a unit selection databases, some are likely to bad, thus quality in a unit selection voices is typical very good most of the time, but occasionally bad.

2. Speech Synthesis Technologies

2.1. Concatenative Synthesis

Although many people believe they need a general TTS voice that can say anything, actually the more you can tailor the voice to it particular application the better it will be. Most high quality speech synthesis voices today offer announcer type voices that are good for giving short informational sentences, mainly because that is what most applications require.

This notion can be explicitly exploited to provide much higher quality synthesis that could be achieved from a general synthesis voice. **Limited Domain** synthesis [3] has been used to refer to systems consisting of only simple pre-recorded prompts, to fully general synthesis voices with some prompts targeted towards the particular application domain. Common template filler type voices like talking clocks or weather are good examples of limited domain voices.

Designing the prompt set for a limited domain voice is crucial to getting a consistent high quality voice. For template voices the basic templates can be identified with the fillers. For examples

> The weather in CITY, STATE, today, DAY is F degrees and the outlook OUTLOOK

The quality of such designed synthesizers can be very good almost all of the time.

Another example of this technique is the CMU Let's Go Bus Information System [4]. A limited domain synthesizer was built to cover not just the set phrases and templates but the 15,000 bus stop names too. The synthesizer is a general synthesizer but it is much better at talking about bus information that about daily news stories.

The point here is that designed voices can provide quality at good as prerecorded prompts, and still offer some level of flexible output.

A second reason for considering designing voices for particular applications is that general voices often sound inappropriate for particular applications. A voice built from news reader speech used in a dialog system may make the user think they are being interviewed on CNN rather than discussing their travel requirements with a travel agent.

2.2. Statistical Parametric Synthesis

Recently a new speech synthesis technique has grown in popularity. Statistical Parametric Speech Synthesis [5] differs from concatenative speech synthesis. Although it is still based on databases of natural speech, instead of instances of particular speech units, models (in their simplest form averages) of units are used to generate speech. The advantage is a better use of the data, thus smaller databases can still produce good results. Because the synthesis output is generated from a model the results are much smoother and have less discontinuities than concatenative speech synthesis techniques. However there are disadvantages too. Because the speech is generated from a parametric model the result is not as sharp as that generated from actual instances of speech. Thus it is sometimes described as sounding "muffled". A second

issue is due to the reconstruction of the signal, as the basic system only represent the spectral part of the signal, the excitation models are limited hence a "buzzy" vocoded output is common, though better parameterization techniques are reducing that (e.g. [6,7]).

The HMM-Generation Synthesis Systems (HTS) [8] is the most famous example of an SPS system. Such techniques have done very well in the annual Blizzard Challenge Speech Synthesis Listening Tests [9].

From a system point of view, SPS also has the advantage of offering more control of the speech output. Unlike unit selection techniques, that require recording sufficient examples of the desired style and content [10], SPS systems can to some extent mix styles. For example emotional speech synthesis is readily possible in an SPS paradigm [11] without having to record large databases of different desired styles.

SPS offers the chance for more flexibility and importantly may offer a route to allow individualized voices without a large amount of careful recording.

2.3. Voice Conversion

Recent technologies have brought the possibility of **voice conversion.** Unlike full voice building, in a voice conversion (sometimes called voice morphing or voice transformation), only a small amount of target speaker data is required which is used to convert an existing larger database.

However in our experiments in trying to find the "best" voice talent to record for the best voice we quickly note that everyone has their own taste. There is one voice that everyone thinks is correct, therefore it is useful to offer a choice of voices. It may be a male user prefers a female voice or a female user prefer a male voice, or vice versa. There seems to be no clear pattern and with speech synthesis voices now sounding like particular people, the users can have very specific unpredictable tastes. "It sounds like my elementary school teacher," may mean they like it (or not).

Early work in voice conversion used a code book technique (e.g. [12]) where a set of acoustic unit types from the target speaker were used. Later more successful techniques, which require a much more computational expensive training algorithm use GMM [13]. In this case a parallel set of sentences is required, i.e. the same sentences from both the source and target speakers. These are first aligned at the frame level, then a GMM is built with the joint source and target frames. At conversion time, the source features are only available and the GMM is used to predict the most likely features give the source features. Features are typically only spectral features such MEL-CEP, thus no excitation transformation occurs, even though it is clear that excitation functions contain speaker identity [14].

3. Synthesis Uses in Educational Systems

3.1. Choosing the right voice

The appropriateness of a voice is critical to its acceptance in an application [15]. If the application is an authoritative source a "newsreader" style may be very appropriate. But if the system is to act like a "buddy" or "peer" it would be better to have a speech style closer to the user. People do assign characters to their talking systems, and if it offers an inappropriate style the user may not be happy.

In offering voice types, there is no single voice that will satisfy all, so it is important to offer a variety and let the user select what they consider to be most appropriate.

Also it is worth considering the non-standard voice. Listeners seem to accept errors in novelty voices more than in the highest quality voice. Basically newsreaders are not supposed to make mistakes while cartoon characters can.

For example, we have built novelty voices like whisper, shouting and "Damien" (a deep "daemon"-like voice). In spite of clear phonetic errors and, especially in the whisper and shouting voice, significant difficulty in understandability, people like the voices. Perhaps because they expect whispering to be harder to understand, they accept the difficulties and blame it on the style rather than the synthesis techniques.

We have also found that non-native voices can be more acceptable that native one. A voice built from my own Scottish accented speech is well accepted by Americans as novel (and even funny), while Scottish listeners do not even immediately recognize it as Scottish. Similar listening tests were also done with a native Chinese speaker, speaking English. It seems that listeners' expectations about errors influence their acceptance of synthetic voices.

3.2. Choosing the right style

The style of the voice also makes a large impact of applicability to an application. As noted "teacher" verses "peer" voices should reflect the intention of the language generation system. When building voices for applications we always explain to the voice talent the intended application and get them to pretend they are delivering lines in that application no matter what the semantic content of the prompts actually are.

Multiple voices may also be an option too. Thus different information may be appropriately delivered in a different voice. For example a "teacher" voice may be used to give general information, a "peer/student" voice may be used for tips and examples, while another voice may be used for anecdotes and real world examples. This variety of voices can enhance a listener's acceptance of the system.

3.3. Having your own voice

It has been suggested that hearing your own voice read things may make it easier to understand. It may just be the novelty of it that makes this useful, and it is clear that not everyone wants to hear their own voice. Speech synthesis technologies have reached the phase where it is reasonable to construct a voice for a particular person. If the language is already supported (see below for when the language is not supported) we offer a number of tools that allow relatively easy construction.

Two basic techniques are possible, full voice construction or just voice conversion. For either we need a set of recorded utterances from the target speaker. These should be recorded as high quality as possible. This both in terms of acoustic record (high quality microphone, studio like recording environment), and speaker accuracy (they must fluent speak the

prompts without error). We have noticed that not everyone has experience in reading prompts fluently. It is important that each prompt is delivered in the same style with the same vocal effort. Thus recording children can be extremely difficult, for children's voices we often use adult actors instead. Although we have had substantial success in building voices with a very large number of different people, with widely various accents, it is clear that it does not always work for all people, even when they appear to read smoothly. However we do note that people get better at delivering prompts with practice.

Full unit selection voices typically require at least 1000 prompts and probably more. Most non-professional voice talent cannot deliver much more than that in a single session. 1000 prompts is about 1 hour of speech, but it will typically take 2 to 3 times that to record it. Statistical Parametric Speech Synthesis techniques like CLUSTERGEN [16] can produce acceptable quality with only 200-300 prompts, and is perhaps more robust to errors.

Voice conversion techniques can work well on many voices with as little as 20-50 prompts. There are two routes for voice conversion. A post synthesis filter may be constructed from a set of 20 utterances. This filter is then applied to an existing synthesizer (for example our standard diphone synthesizer) to get speaker identity in synthesis. This works fairly well depending on the target speaker's voice. This technique is the quickest and easiest to carry out.

The second technique is to use the conversion model to convert a larger database then build a complete voice. This will typically give better results but requires more work.

3.4. Having your own language

Tools like FestVox [17] have existed, for some time, allowing an interested party to build speech synthesis voices in any language. FestVox has been used for at least 40 different languages through out the world.

The tasks involved in building synthesis support in a new language are becoming better defined, but there certainly still somewhat of an art. Often there are phenomena in a new language which may require special processing. For example, no spaces between words (e.g. Chinese or Thai), no written vowels (e.g. Arabic or Farsi), no stress markings in orthography (e.g. Russian or English).

The basic tasks involved in constructs a synthesizer in a new language are:

- Collect a large amount of example text: for finding prompts, and lexical frequencies.
- Define a phoneme set for the language.
- Construct a pronunciation lexicon and letter to sound rules for unknown words.
- Select a set of prompts to record that cover the phonetic and prosodic variation in the language.
- Record the prompts.
- Build a synthesizer by labeling the prompts and constructing acoustic and prosodic models.
- Define text analysis rules: for expanding numbers, symbols etc.
- Evaluate, tune and test resulting synthesizer.

The amount of work required for each stage may vary from language to language depending on its inherent complexity, but also varies based on the required coverage. If the type of text to be synthesized is mostly a closed class vocabulary with no significant use of symbols, then lexicon construction and text analysis may be much simpler.

More recently we have developed a web-based tool aimed at constructing both recognition and synthesis models in new languages. The SPICE project [18] is aimed at non-speech experts. It clearly leads the developer through the necessary stages in building recognition and acoustic models. It has been used successfully for some 12 different languages and the feedback from that development is being incorporated into the system [19].

The SPICE system builds a CLUSTERGEN parametric voice, from a prompt list that is designed as part of the build process. The prompts consist of "nice" sentences of around 5-15 words using only high frequency words thus making the sentence easy to say and less likely for the voice talent to make errors.

3.5. Having your own voice in any language

It has been suggested that hearing your own voice in a new language will help you be able to speak that language more clearly. Although this may not work for all speakers it does seem like a useful capability. This goal of cross language voice conversion is a current hot topic. One of the driving forces, for this work, is speech-to-speech translation where it would be appropriate for the output translated voice to sound like the source speaker.

Standard voice conversion techniques will not work for this, as the speech from the target speaker speaking in L1 will not contain phonetic variation required for L2.

A speaker's voice effectively contains two components the **language component**: how phonemes are realized in the particular language/dialect; and a **speaker component**: how the speaker themselves realize these phones: in their particular idiolect. Voice conversion techniques conflate these two components.

One proposed solution to this is to attempt to model the different between the languages separately [20]. Suppose we have a bi-lingual speaker for the two languages of interest. We collect data from the speaker in the two languages. Unlike the normal case of building a transformation mode, these sentences cannot be aligned as they are in a different language. Thus an alignment first based on phonetic mapping is required before training a GMM model for spectral conversion. This current technique however is still limited to finding appropriate bilingual speakers to bootstrap a system.

4. Conclusions

This paper has present the current state-of-the-art in speech synthesis technology, highlight the areas that may be useful for educational systems. The important message is that quality of speech synthesis output has drastically improved, but it is still worthwhile spending time selecting the right quality for the

right application. Recording prompts in itself is time consuming, and not trivial to get right, thus using standard high quality synthesizers may provided more consist quality and of course be much easier to update.

Selecting the right voice can be critical in giving the right style for the application, and this may be more important than the technical quality of the voice.

In language education systems, phonetic quality may be more important than overall fluency, thus targeted recording of a "golden voice" may be more reliable, than a synthesizer, though of course will be more restrictive.

The final piece of technology that, although is still developing, if ready to be used at least in experiments is voice conversion. As it is not too hard to make systems begin to sound like the user of the system, we could set up experiments to see if this helps.

True cross-language voice conversion is still someway off but will be available soon.

5. References

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

[2] Olive, J., Greenwood, A. and Coleman, J. "Acoustics of American English Speech: A Dynamic Approach", Springer Verlag, 1993.

[3] Black, A. and Lenzo, K. "Limited Domain Speech Synthesis", ICSLP 2000, Beijing, China, 2000.

[4] Raux, A., Langner, B., Bohus, D., Black, A. and Eskenazi, M. "Let's Go Public! Taking a Spoken Dialog System to the Real World", Interspeech 2005, Lisbon, Portugal, 2005.

[5] Special Session on Statistical Parametric Speech Synthesis, at ICASSP 2007, Honolulu, Hawaii 2007.

[6] Kawahara, H., Masuda-Katsuse, I., and Cheveigne, A. "Restructuring Speech Representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: possible role of a repetitive structure in sounds", Speech Communication, 27, 187-207. 1999.

[7] Stylianou, Y. "Concatenative Synthesis using a Harmonic plus Noise model" 3rd ESCA Speech Synthesis Workshop, Jenolan Caves, NSW, Australia, 1998.

[8] Tokuda, K, Zen, Heiga and HTS Working Group, "HTS: HMM-based speech synthesis system", Nagoya Institute of Technology, http://hts.ics.nitech.ac.jp 2001.

[9] Bennett, C. "Large Scale Evaluation of Corpus-Based Synthesizers: Results and Lessons from the Blizzard Challenge 2005", Interspeech 2005, Lisbon, Portugal, 2005. http://festvox.org/blizzard.

[10] Eide, E., Aaron, A., Bakis, R., Hamza, W., Picheny, M. and Pitrelli, J. "A corpus-based approach to <AHEM> Expressive Speech Synthesis", 5th ISCA Speech Synthesis Workshop, Pittsburgh, PA, 2004.

[11] Bulut, M., Lee, S., and Narayanan, S. "A Statistical Approach for modeling prosody features using POS tags for Emotional Speech Synthesis", ICASSP 2007, Homolulu, Hawaii, 2007.

[12] Abe, M., Shikano, K. and Kuwabara, H. "Voice conversion through vector quantization", J. Acoust. Soc. Jpn. (E), 11 71-76, 1990.

[13] Stylianou, Y., Cappe, O., and Moulines, E. "Statistical Methods for voice quality transformation." Eurospeech 1995, 447-450, Madrid, Spain, 1995.

[14] Kain, A. "High Resolution Voice Transformation", PhD Thesis, Oregon Graduate Institute, 2001.

[15] Nass, C., and Brave, S. "Wired for Speech: How Voice Activates and Advances the Human Computer Relationship", MIT Press, 2005.

[16] Black, A., "CLUSTERGEN: A statistical parametric synthesizer using trajectory modeling", Interspeech 2006, Pittsburgh, PA. 2006.

[17] Black, A. and Lenzo, K. "FestVox: Building Synthetic Voices" http://festvox.org 2000.

[18] Schultz, T. and Black, A. "Challenges with Rapid Adaptation of Speech Translation Systems to New Language Pairs", ICASSP 2006, Toulouse, France, 2006.

[19] Schultz, T., Black, A., Badaskar, S., Hornyak, M., Kominek, J. "SPICE: Web-based tools for rapid language adaptation in speech processing systems" Interspeech 2007, Antwerp, Belgium, 2007.

[20] Mashimo, M., Toda, T., Kawanami,H., Shikano, K., and Campbell, N., "Cross-language voice conversion evaluation using bilingual speaker databases", IPSJ, 43(7), pp 2177-2185, 2002.