Proceedings of the Institute of Acoustics

SPEECH SYNTHESIS

A. P. Breen

BT Laboratories, MLB 3/44, Martlesham Heath, Ipswich, Suffolk, IP5 3RE.

1. INTRODUCTION

The desire to produce a speaking machine has been an enduring pastime for mankind. Early research into the field focused on mechanical models of the vocal apparatus[1]. With the coming of the telecommunications age, the fundamental approach to synthesis changed from mimicking to the production system, to modelling the acoustic signal[2].

It's unclear what motivated early attempts at speech synthesis, possibly simple intellectual curiosity, or there may have been an underlying assumption that the ability to produce synthetic speech implied the ability to portray intelligence. The initial impetus for developing synthesis systems in telecoms was as a compression. With hindsight, this application now seems poorly suited to speech synthesis. Academics have been interested in speech synthesisers as a means of investigating aspects of speech perception. Such researchers are not in the business of developing "complete" systems, which are the subject of this paper.

Up until comparatively recently, a great deal of effort was directed at the generation of synthetic speech from some form of low level parametric interface. Some notable complete text to speech systems were produced [3][4], but such systems failed to make a significant impact as products.

Speech synthesis was given a boost in the eighties and early nineties with the marked reduction in cost of computer memory and the development of efficient signal processing techniques such as PSOLA. These signal processing algorithms enabled researchers to efficiently modify the pitch and duration of a speech signal without introducing unacceptable distortion. This in conjunction with efficient unit selection and joining algorithms provided text to speech systems with a synthesisers capable of a naturalness previously unachievable by formant based systems, except through hand crafted copy synthesis.

Speech researchers quickly realised that the improved segmental quality of speech synthesis simply served to emphasie a lack of understanding of the prosodic aspects of speech synthesis. Much of the improvements in the quality of speech synthesis are in fact attributable to a greater trend in modelling ignorance through data driven analysis and encoding systems, than through any significant advance in our understanding of the process of speech production.

This paper will not provide an overview of text to speech systems, there are a number of publications which perform this task[5][6]. Instead, I hope to highlight some of the major issues facing today's system developers. These issues fall into a number of categories some technical, some practical and some philosophical. They represent topics of immediate importance to me and, I believe, of general relevance to researchers and developers working in the field. This paper is primarily concerned with limitations! Limitations of the current technology, our level of understanding and in our appreciation of what is achievable with this technology. I hope that the discussions provided below will go some way to explaining why speech synthesis is still a immature technology.

2. TIME-FREQUENCY MODIFICATION ALGORITHMS

Pitch Synchronous overlap Add (PSOLA) and its derivatives [7] form the back bone of many synthesisers used in today's text to speech systems. As stated in the introduction, the development of algorithms such as this played a significant part in changing the direction of research on text to speech synthesis. For these reasons its worth spending some time considering its benefits and short falls. The two main benefits of time domain time-frequency modification algorithms are that they are efficient and capable of modifying the pitch and duration of a speech signal without introducing significant audible distortion. However the degree of distortion introduced depends to a large extent on the properties of the original speech signal and on the size of modification requested. Typically such methods work well for frequency modifications of less than an octave, and duration modifications less than twice the original length. The three major limitations of the standard overlap add methods are listed below:

1. Generally they do not work well with large abrupt changes in pitch. This is predominantly due to their inability to modify the underlying spectral properties of the speech signal.
2. They do not work well with breathy or creaky voice. In other words, when the speech signal has either a very long or very short open phase in the larynx cycle.
3. They do not work well with unvoiced sounds such as fricatives and other complex sounds such as africates and plosives.

Alternative methods of time-frequency modification [7][8] have been suggested, whilst these are all computationally more expensive than simple time domain overlap-add they provide greater control over the spectral properties of the speech signal, and offer better performance in the modification of unvoiced sound such as fricatives. The disadvantages of such methods are; that they are computationally more expensive, not as robust as the simpler methods and produce speech of a slightly lower overall quality. All these deficiencies stem from the fact that the more powerful methods rely on an underlying model of speech production.

The spectral properties of the speech signal are affected by a number of events, simply changing the pitch range of speech forces changes in the quality of the voice, but speaking style is equally, if not more, significant.
There are two approaches open to researchers wishing to improve the speech quality produced by such algorithms:

1. The next generation of algorithms must model the effect pitch has on the spectral properties of the speech signal, they must take more account of the characteristics of the original signal and finally have some capacity for generating information not present in the original signal.
2. An alternative approach is to recognise that any manipulation of the speech signal will result in some degree of degradation, and thus employ a unit inventory and selection process which selects units on the basis of how closely the properties of the stored phone match the desired pitch and duration. This second approach leads on to the next problem area, unit selection.

3. UNIT SELECTION

The last section highlighted the fact that the current generation of time-frequency modification algorithms are insensitive to the segmental quality of the underlying concatenated speech units. Today's speech synthesisers work well because they synthesise a particular style of speech. Attempts to change style falter because they break underlying assumptions made during the generation of the unit database, the choice of unit, the selection criteria and the choice of signal modification algorithm. This section provides a brief overview of commonly used methods of unit selection. The process used within the Laureate text to speech system provides a specific example, but much of the discussion is generally applicable.

The primary goal of a unit selection process is to make best use of the available information contained within the finite data set. However, other practical factors such as processing time and the desire to minimise the number of unit discontinuities also play an important part in the design of a system. The type of unit to be stored in the inventory of sounds is important as this choice greatly effects the recording process, the size of the inventory and the method of unit selection. The commonest forms of unit are the diphone and the triphone. Typically these units are recorded within a carefully designed set of carrier phrases. An alternative to the fixed inventory approach, is to dynamically select segments of speech from a very large corpus. Between these two approaches lie a number of methods that attempt to optimise the design of the database based on specific linguistic criteria e.g. syllable structure. Inevitably, the nature of the recorded database will depend upon the underlying assumptions of the unit selection process. An additional consideration for many researchers is the size of the database. Practical limitations on size significantly affect the type of database recorded and hence the sophistication of the unit selection process employed.
The speech database design used in Laureate has similarities with the approaches mentioned above. Large speech databases are considered difficult to maintain and even more difficult to annotate reliably. As a result a fairly simple adequacy criterion is employed. The speech database is designed such that it contains at least one instance of every diphone permitted by the pronunciation model. However, it differs from many such databases in that it is not composed from diphones embedded within a set of carrier phases. Instead, the database consists of phonetically rich passages. Diphone coverage only represents a minimum adequacy criteribn. The unit selection process is not restricted to selecting diphones but is free to select N-phone units. In practice units are restricted to a maximum of three phones drawn from a five phone context. This process is described in detail in [9][10].

In reality the choice of unit type is not the most significant aspect of the design. The two most important factors are:
1. The linguistically significant features used in the selection process.
2. The degree of coverage.

The first point highlights the fact that it does not matter what size of database is used or what flavour of unit selected, if the database has not been annotated in a meaningful way. The phoneme is commonly used as an appropriate label. However, phonemes, while pragmatically useful, are a relatively poor encoders of linguistic information. Systems which rely on solely phonemic identity in the selection process are drastically under utilising the recorded database.
The second point highlights the fact that without sufficient coverage, the most sophisticated labelling system will fail to provide natural synthesis. Coverage should not be confused with size, as a large database does not guarantee adequate coverage, care must be taken to ensure that the database contains as many discriminating alternatives as possible.
Provided care is taken in the generation of the database, and an appropriate labelling system adopted, high quality synthesis is almost guaranteed. There is of course a catch! Which is this. The selection process is only as good as the information contained within the annotated speech database and the information supplied by the rest of the speech synthesis system. A sophisticated annotation strategy assumes that the remaining components of the system are capable of providing an equally rich description of the required sounds.

The introduction suggested that to date we still have an incomplete picture of how speech is generated, and what factors are influential in its production and perception. The degree of interaction between purely structural aspects of the speech signal and those relating to paralinguistic factors such as emotion, style and discourse are still unclear. This lack of understanding is clearly audible in synthetic speech. There are basically two paths taken by researchers attempting to deal with this problem. Those who consider data driven approaches, and those who consider theory driven approaches. Both paths have significant drawbacks. Data driven methods have proved effective in certain areas, such as duration prediction, but often suffer from a lack of adequate training material and the inability to generalise. In general data driven approaches assume that structural information useful to speech synthesis is available from a surface analysis. However, phonological models such as autosegmental phonology would suggest that much of the relevant structural information is not present in the speech signal. Theory driven approaches fare

no better. Rule based systems have been discredited and linguistic theories tend to be incomplete, fragile and difficult to compute. As an example, a great deal of research into intonation is descriptive [11] rather than predictive. From the above the discussion, it should be clear that there are significant problems in attempting to model any style let alone attempting emotive speech styles.

Comparatively little research has been conducted into speaking styles within the synthesis process, yet as stated above, inventory based systems are particularly sensitive to the manner of speech recorded for the database. Investigations of synthesis style have centred predominantly in the area of emotional synthesis [12], where researchers have concentrated on simulating basic emotions such as fear, happiness, sadness and boredom.
In designing the speech database for the Laureate system for example, care has been taken to ensure that the recorded speech does not exhibit any strong speaking style, but maintains a neutral, placid quality. Where feasible and appropriate, specific speaking style effects may be imposed on the speech as part of the post selection synthesis process, but as discussed in the previous section the degree of flexibility allowed by the current generation of overlap-add methods is limited. One way of alleviating this problem, also suggested in the last section, is to include pitch and duration information in the labelling system. Apart from the obvious consequence of drastically increasing the size of a database, this approach is complicated by the fact that extracting reliable pitch estimates from speech data is a non-trivial process.

4. WHO WANTS IT ANYWAY?

While there are clearly a number of applications in telecommunications where the current generation of TTS would be useful [13], the fact remains, that the full potential of synthesis is far from being realised. Most application developers are staying with recorded speech despite the fact that this involves a time consuming and costly data recording process. The reason is simple, application developers do not believe that the quality of synthetic speech is good enough. The vast majority of applications require synthetic speech to exhibit some specific emotion or style. In other words, the very attributes that synthesis systems find hard to model. But are application developers asking for the impossible? Should researchers be attempting to build text to speech systems? Which as the name suggests take as their input plain text. Plain text is a poor encoder of information, which means that TTS systems are forced to undertake some form of shallow text analysis.

The majority of typographical ambiguities found in unrestricted text are handled by a process known as text normalisation, while gross cases may be handled using a limited set of escape sequences. Plain text however does not contain sufficient information to correctly resolve many types of ambiguity. As a result, a large number of the decisions made by the text normalisation process are in effect arbitrary. This problem is not restricted to the process of text normalisation, but seriously effects all aspects of the speech synthesis process. To illustrate this point, consider the following simple example. Imagine that as part of some dialogue, an automated system has generated the question given below:

Did you say fifty pence?

The meaning of this question changes with word emphasis. For example, if emphasis was placed on the word *pence*, the system is asking the user to confirm that the amount was in pence rather than pounds, whereas, if emphasis was placed on the word *fifty*, the system is asking the user to confirm that the number was fifty as opposed to some other amount. By default, with only plain text to work with, a text to speech systems would typically place the emphasis on "pence".

A TTS system may be asked to convert text from a news article, a poem, a discourse, or even a train timetable. Typically, little or no pre-processing will be performed on the text. Is it any wonder then, that the quality of speech produced by such systems is so stilted?

TTS researchers are faced with two choices. Either to increase the level of linguistic analysis conducted on text within the TTS system, or to encourage users to extend the type of information presented to a speech synthesis system.

If the first choice is taken, TTS systems will be forced to balloon in size, effectively taking on the majority of the tasks involved in the interpretation and presentation of information. This is an unpalatable choice for many synthesis researchers, as linguistic analysis, while being a necessary requirement for the production of synthetic speech, is by no means a sufficient requirement. Knowing what to say is not the same as knowing how to say it. Researchers in speech synthesis still have a lot to learn about the human production system. Unfortunately working with plain text is effectively stifling research into production.

If the second choice is taken, researchers are obliged to investigate effective methods of encoding linguistic and para-linguistic information, which in itself is currently an ill defined and complex task.

5. LEVELS OF REPRESENTATION

What is the future of text to speech synthesis? Speech synthesis systems have improved, telecommunications is still a major driving force in the development of speech synthesis technology. Synthesis is no longer seen simply as a means of efficiently compressing information, but potentially, as an important part of advanced telecommunications services. The explosion of the personal computer market has opened up an alternative path for speech synthesis systems. Particularly as it is starting to appear as a basic component of a computers operating system. The appearance of interfaces such as Microsoft's SDK and the JAVA speech API offer a much wider choice of systems for applications developers, and potential applications for TTS. It remains to be seen if this increase in the accessibility will eventually lead to an increase in its use within serious applications or consign it to the realms of technological toys.

Speech researchers have recognised that, the current model of text to speech synthesis must be changed. Proposed mark-up languages such as SSML and SABLE [14][15][16] are a first attempt at addressing the problem. The defacto standardisation of interfaces such as those outlined in SDK are making it more difficult for system developers to change the behaviour of their systems, which in turn are making mark-up proposals such as SABLE more attractive. But are these languages providing application developers with the right information? It seems to me that the vast majority of application developers want a mark-up system which provides a high level abstract description of speaking style, emotion, and discourse acts rather than a set of comparatively low level acoustic controls. If we as speech researchers are unable to produce natural sounding speech with the much richer set of controls available at the system level why should others be able to achieve it using the more limited set of controls provided by an applications interface layer?

6. CONCLUSION

This review paper has introduced a number of issues facing speech synthesis system developers. The topics discussed do not represent a complete or definitive list. Other researchers in the field are certain to disagree with some or all of the points I have outlined above. The basic fact still remains, that speech synthesisers are not producing the type of speech wanted by most application developers. It is not simply a case of intelligibility or even of how "human" the synthetic speech sounds, although these are still significant factors. The current generation of synthesisers produced bored, disinterested sounding speech. Speech which is lacking in any emotion and warmth. It is predominantly for this reason rather than any other that application developers are slow to take up this technology. There is little likelihood of providing this sort of performance from within the existing paradigm of text to speech synthesis. Application developers will need to provide explicit information about mood and intention, while speech researchers will need to develop algorithms which are capable of realising such abstract concepts.

7. REFERENCES

[1] Flanagan, A. J., Voices of men and machines, J. Acoust. Soc. Am., Vol. 63, 1972.
[2] Dudley, H., Tarnoczy, T.H., The speaking machine of Wolfgang Von Kemplen, J. Acoust. Soc. Am., Vol. 22, No.1, 1950.
[3] Holmes, J.N., Mattingly, I.G., Shearme, J.N., Speech synthesis by rule, Language and speech, No. 7, 1964.
[4] Allen, J., Hunnicutt, M. S., Klatt, D., *From text to speech: The MITalk system*, Cambridge University Press, 1987.
[5] Klatt, D.H., Review of text-to-speech conversion for English, J. Acoust. Soc. Am., Vol. 82, No. 3, 1987.
[6] Edgington. M., Lowry. A., Jackson. P., Breen A. P., Minnis. S. Overview of Current Text-to-Speech Techniques: Part I - II, BT Technol. J.:14:1., 1996.
[7] Syrdal, Stylianou et al.,TD-PSOLA versus Harmonic Plus Noise Model in Diphone Based Speech Synthesis, ICASSP-98.
[8] Violaro, Boeffard,A hybrid model for text-to-speech synthesis, IEEE Trans. Speech & audio Vol. 6, No 3, 1998.
[9] Breen, A. P., Jackson, P.,Non-uniform unit selection within BT's Laureate text to speech system, To appear in Proc. ICSLP 98, Sydney, Australia, 1998.
[10] Breen, A. P., Jackson, P.,A phonologically motivated method of selecting non-uniform units, To appear in Proc. ESCA / COCOSDA 3rd International Workshop on Speech Synthesis, Jenolan Caves, Australia,1998.
[11] Working Papers 41., ESCA Workshop on Prosody, Lund, Sweden, 1993.
[12] Murray, I.R., Arnott, J. L., "Synthesising emotion in speech: is it time to get excited?", Proc. ICSLP 96, 1996.
[13] Page, J. H., Breen, A. P., "The Laureate text to speech system – architecture and applications", *Speech Technology for Telecommunications*, Chapman and Hall, 1998.
[14] Slott, J. M. A Generalised Platform and Markup Language for Text to Speech Synthesis. Ph.D. diss., Dept Electrical Engineering and Computer Science, MIT. 1996.
[15] Taylor P, Isard A. "SSML: A Speech Synthesis Markup Language", Speech Communication:21:123-133. 1997.
[16] SABLE: A synthesis markup language (version 2.0), http://www.bell-labs.com/project/tts/sable.html.