Towards Greater Naturalness:
Future Directions of Research in Speech Synthesis

Eric Keller

Laboratoire d'analyse informatique de la parole (LAIP)

IMM-Lettres, University of Lausanne, 1015 Lausanne, Switzerland

Eric.Keller@imm.unil.ch

In the past ten years, many speech synthesis systems have shown remarkable improvements in quality. Instead of monotonous, incoherent and mechanical-sounding speech utterances, these systems produce output that sounds relatively close to human speech. To the ear, two elements contributing to the improvement stand out, improvements in signal quality on the one hand, and improvements in coherence and naturalness on the other. These elements reflect in fact two major technological changes. The improvements in signal quality of good contemporary systems are mainly due to the use and improved control over concatenative speech technology, while the greater coherence and naturalness of synthetic speech is primarily a function of much improved prosodic modelling.

However, as good as some of the best systems sound today, few listeners are fooled into believing that they hear human speakers. Even when the simulation is very good, it is still not perfect -- no matter how one wishes to turn the issue. Given the massive research and financial investment from which speech synthesis has profited over the years, this general observation evokes some exasperation. The holy grail of 'true naturalness' in synthetic speech seems so near, and yet so elusive. What in the world could still be missing?

As so often, the answer is complex. The present volume introduces and discusses a great variety of issues affecting naturalness in synthetic speech. In fact, at one level or another, it is probably true that most research in speech synthesis today deals with this very issue. To lead off the discussion, this article presents a personal view of recent encouraging developments and continued frustrating limitations of current systems. This in turn will lead to a description of the research challenges to be confronted over the coming years.

The Current Status

Signal Quality and the Move to Time-Domain Concatenative Speech Synthesis

The first generation of speech synthesis devices capable of unlimited speech (Klatt-Talk, DEC-Talk, or early InfoVox synthesisers) used a technology called 'formant synthesis' (Klatt, 1989; Klatt & Klatt, 1990; Styger & Keller, 1994). While speech produced by formant synthesis produced the classic 'robotic' style of speech, formant synthesis was also a remarkable technological development that has had some long-lasting effects. In this approach, *voiced* speech sounds are created much as one would create a sculpture from stone or wood: a complex waveform of harmonic frequencies is created first, and 'the parts that are too much', *i.e.* non-formant frequencies, are suppressed by filtering. For *unvoiced* or *partially voiced* sounds, various types of noise are created, or are mixed in with the voiced signal. In formant synthesis, speech sounds are thus created entirely from equations. Although obviously modelled on actual speakers, a formant synthesiser is not tied to a single voice. It can be induced to produce a great variety of voices (male, female, young, old, hoarse, etc).

However this approach also posed several difficulties, the main one being that of excessive complexity. Although theoretically capable of producing close to human-like speech under the best of circumstances (YorkTalk a-c, CD-ROM), these devices must be fed a complex and coherent set of parameters every 2-10 ms. Speech degrades rapidly if the coherence between the parameters is disrupted. Some coherence

constraints are given by mathematical relations resulting from vocal tract size relationships, and can be enforced automatically via algorithms developed by Stevens and his colleagues (Stevens, 1998). But others are language- and speaker-specific and are more difficult to identify, implement, and enforce automatically. For this reason, really good-sounding synthetic speech has to my knowledge not ever been produced entirely automatically with formant synthesis.

The apparent solution for these problems has been the general transition to 'time-domain concatenative speech synthesis' (TD-synthesis for short). In this approach, large databases are collected, and constituent speech portions (segments, syllables, words, and phrases) are identified. During the synthesis phase, designated signal portions (diphones, polyphones, or even whole phrases [1]) are retrieved from the database according to phonological selection criteria ('unit selection'), chained together ('concatenation'), and modified for timing and melody ('prosodic modification'). Because such speech portions are basically stored and minimally modified segments of human speech, TD-generated speech consists by definition only of possible human speech sounds, which in addition preserve the personal characteristics of a specific speaker. This accounts, by and large, for the improved signal quality of current TD speech synthesis.

Prosodic Quality and the Move to Stochastic Models

The second major factor in recent improvements of speech synthesis quality has been the refinement of prosodic models (see 'State-of-the-Art Summary...' by Monaghan, this volume, plus further contributions found in the prosody section of this volume). Such models tend to fall into two categories, predominantly linguistic and predominantly empirical-statistic ('stochastic'). For many languages, early linguistically-inspired models did not furnish satisfactory results, since they were incapable of providing credible predictive timing schemas or the full texture of a melodic line. The reasons of these insufficiencies are complex. Our own writings have criticised the exclusive dependence on phonosyntax for the prediction of major and minor phrase boundaries, the difficulty of recreating specific Hertz values for the fundamental frequency ('melody', abbr. F0) on the basis of distinctive features, and the strong dependence on the notion of 'accent' in languages like French where accents are not reliably defined (Zellner, 1996, 1998; Keller et al., 1997).

As a consequence of these inadequacies, so-called 'stochastic' models have moved into the dominant position among high-quality speech synthesis devices. These generally implement either an array or a tree structure of predictive parameters and derive statistical predictors for timing and F0 from extensive database material. The prediction parameters do not change a great deal from language to language. They generally concern the position in the syllable, word and phrase, the sounds making up a syllable, the preceding and following sounds, and the syntactic and lexical status of the word (e.g., Keller and Zellner, 1996; Zellner Keller & Keller, in press). Models diverge primarily with respect to the quantitative approach employed (*e.g.*, artificial neural network, classification and regression tree, sum-of-products model, general linear model (Campbell, 1992; Riley, 1992; Keller & Zellner, 1996; Zellner Keller & Keller, this volume), and the logic underlying the tree structure.

While stochastic models have brought remarkable improvements in the refinement of control over prosodic parametres, they have their own limitations and failures. One notable limit is rooted in the 'sparse data problem' (van Santen & Shih, 2000). That is, some of the predictive parameters occur a great deal less frequently than others, which makes it difficult to gather enough material to estimate their influence in an overall predictive scheme. Consequently a predicted melodic or timing parameter may be 'quite out of line' every once in awhile. A second facet of the same sparse data problem is seen in parameter interactions. While the effects of most predictive parameters is approximatively cumulative, a few parameter combinations show unusually strong interaction effects. These are often difficult to estimate, since the contributing parameters are so rare and enter into interactions even less frequently. On the whole, 'sparse data' problems are solved in either a 'brute force' approach (gather more data, *much* more), by careful binning of data (e.g., establish sound groups, rather than model sounds individually), and/or by resorting to a set of supplementary rules that 'fix' some of the more obvious errors induced by stochastic modelling.

A further notable limit of stochastic models is their averaging tendency, well illustrated by the problem of modelling F0 at the end of sentences. In many languages, questions can end on either a higher or a lower F0 value than that used in a declarative sentence (as in "is *that* what you mean?"). If high-F0 sentences are not rigorously, perhaps manually, separated from low-F0 sentences, the resulting statistical predictor value will tend towards a mid-F0 value, which is obviously wrong. A fairly obvious example was chosen here, but the problem is pervasive and must be guarded against throughout the modelling effort.

The Contribution of Timing

Another important contributor to greater prosodic quality has been the improvement of the prediction of timing. Whereas early timing models were based on simple average values for different types of phonetic segments, current synthesis systems tend to resort to fairly complex stochastic modelling of multiple levels of timing control (Campbell, 1992a, 1992b; Keller & Zellner, 1996, Zellner 1996, 1998a, b).

Developing timing control that is precise as well as adequate to all possible speech conditions is rather challenging. In our own adjustments of timing in a French synthesis system, we have found that changes in certain vowel durations as small as 2% can induce audible improvements or degradations in sound quality, particularly when judged over longer passages. Further notable improvements in the perceptual quality of prosody can be obtained by a careful analysis of links between timing and F0. Prosody only sounds 'just right' when F0 peaks occur at expected places in the vowel. Also of importance is the order and degree of interaction that is modelled between timing and F0. Although the question of whether timing or F0 modelling should come first has apparently never been investigated systematically, our own experiments have suggested that timing feeding into F0 gives considerably better results than the inverse (Zellner, 1998; Keller et al., 1997; Siebenhaar et al., this volume). This modelling arrangement permits timing to influence a number of F0 parameters, including F0 peak width in slow and fast speech modes.

Upstream, timing is strongly influenced by phrasing, or the way an utterance is broken up into groups of words. Most traditional speech synthesis devices were primarily guided by phonosyntactic principles in this respect. However in our laboratory, we have found that psycholinguistically-driven dependency trees oriented towards actual human speech behaviour seem to perform better in timing than dependency trees derived from phonosyntactic principles (Zellner, 1997). That is, our timing improves if we attempt to model the way speakers tend to group words in their real-time speech behaviour. In our modelling of French timing, a relatively simple, psycholinguistically-motivated phrasing ('chunking') principle has turned out to be a credible predictor of temporal structures even when varying speech rate (Keller et al., 1993; Keller & Zellner, 1996). Recent research has shown that this is not a peculiarity of our work on French, because similar results have also been obtained with German (Siebenhaar et al., this volume).

To sum up recent developments in signal quality and prosodic modelling, it can be said that a typical contemporary high-quality system tends to be a TD-synthesis system incorporating a series of fairly sophisticated stochastic models for timing and melody, and less frequently, one for amplitude. Not surprisingly, better quality has led to a much wider use of speech synthesis, which is illustrated in the next section.

Uses for High-Quality Speech Synthesis

Given the robot-like quality of early forms of speech synthesis, the traditional application for speech synthesis has been the simulation of a 'serious and responsible speaker' in various virtual environments (e.g., a reader for the visually handicapped, for remote reading of email, product descriptions, weather reports, stock market quotations, etc.). However the quality of today's best synthesis systems broadens the possible applications of this technology. With sufficient naturalness, one can imagine automated news readers in virtual radio stations, salesmen in virtual stores, or speakers of extinct and recreated languages.

High-quality synthesis systems can also be used in places that were not considered before, such as assisting language teachers in certain language learning exercises. Passages can be presented as

frequently as desired, and sound examples can be made up that could *not* be produced by a human being (e.g., speech with intonation, but no rhythm), permitting the training of prosodic and articulatory competence. Speech synthesisers can slow down stretches of speech to ease familiarisation and articulatory training with novel sound sequences (LAIPTTS a, b, CD-ROM [2]). Advanced learners can experiment with the accelerated reproduction speeds used by the visually handicapped for scanning texts (LAIPTTS c, d, CD-ROM). Another obvious second-language application area is listening comprehension, where a speech synthesis system acts as an 'indefatigable substitute native speaker' available 24 hours a day, anywhere in the world.

A high-quality speech synthesis could further be used for literacy training. Since illiteracy has stigmatising status in our societies, a computer can profit from the fact that it is *not* a human, and is thus likely to be perceived as non-judgemental and neutral by learners. In addition, speech synthesis could become a useful tool for linguistic and psycholinguistic experimentation. Knowledge from selected and diverse levels (phonetic, phonological, prosodic, lexical, etc.) can be simulated to verify the relevance of type of knowledge individually and interactively. Already now, speech synthesis systems can be used to experiment with rhythm and pitch patterns, the placement of major and minor phrase boundaries, and typical phonological patterns in a language (LAIPTTS e, f, i-l, CD-ROM). Finally, speech synthesis increasingly serves as a computer tool. Like dictionaries, grammars (correctors) and translation systems, speech synthesisers are finding a natural place on computers. Particularly when the language competence of a synthesis system begins to outstrip that of some of the better second language users, such systems become useful new adjunct tools.

Limits of Current Systems

But rising expectations induced by a wider use of improved speech synthesis systems also serve to illustrate the failings and limitations of contemporary systems. Current top systems for the world's major languages not only tend to make some glaring errors, they are also severely limited with respect to styles of speech and number of voices. Typical contemporary systems offer perhaps a few voices, and they produce essentially a single style of speech (usually a neutral-sounding 'news-reading style'). Contrast that with a typical human community of speakers, which incorporates an enormous variety of voices and a considerable gamut of distinct speech styles, appropriate to the innumerable facets of human language interaction. While errors can ultimately be eliminated by better programming and the marking up of input text, insufficiencies in voice and style variety are much harder problems to solve.

This is best illustrated with a concrete example. When changing speech style, speakers tend to change timing. Since many timing changes are non-linear, they cannot be easily predicted from current models. Our own timing model for French, for example, is based on laboratory recordings of a native speaker of French, reading a long series of French sentences -- in excess of 10 000 manually measured segments. Speech driven by this model is credible and can be useful for a variety of purposes. However, this timing style is quite different from that of a well-known French newscaster recorded in an actual TV newscast. Sound example TV_BerlinOrig.wav is a short portion taken from a French TV newscast of January 1998, and LAIPTTS h, CD-ROM, illustrates the reading of the same text with our speech synthesis system. Analysis of the example showed that the two renderings differ primarily with respect to timing, and that the newscaster's temporal structure could not be easily derived from our timing model [3]. Consequently in order to produce a timing model for this newscaster, a large portion of the study underlying the original timing model would probably have to be redone (*i.e.*, another 10 000 segments to measure, and another statistical model to build).

This raises the question of how many *speech styles* are required in the absolute. A consideration of the most common style-determining factors indicates that it must be quite a few (Table 1). The total derived from this list is 180 (4*5*3*3) theoretically possible styles. It is true that the Table 1 is only indicative: there is as yet no unanimity on the definition of 'style of speech' or its 'active parameters' (see the discussion of this issue by Terken, this volume). Also some styles could probably be modeled as variants of other styles, and some parameter combinations are impossible or unlikely (a spelled, commanding presentation of questions, for example). While some initial steps towards expanded styles

of speech are currently being pioneered (see the articles in this volume in the 'styles-of-speech' section), it remains true that only very few of all possible human speech styles are supported by current speech synthesis systems.

Table 1. Theoretically Possible Styles of Speech

Parameter	Instantiations	N
Speech rate	spelled, deliberate, normal, fast	4
Type of speech	spontaneous, prepared oral, command, dialogue, multilogue, reading	5
Material-related	continuous text, lists, questions, (perhaps more)	3
Dialect	(dependent on language and grain of analysis)	3

Emotional and expressive speech constitutes another evident gap for current systems, despite a considerable theoretical effort currently directed at the question (Ní Chasaide and Gobl, this volume; Zei and Archinard, this volume; ISCA workshop, www.qub.ac.uk/en/isca/index.htm). The lack of general availability of emotional variables prevents systems from being put to use in animation, automatic dubbing, virtual theatre, etc. It may be asked how many voices would theoretically be desirable. Table 2 shows a list of factors that are known to, or can conceivably influence, voice quality. Again, this list is likely incomplete and not all theoretical combinations are possible (it is difficult to conceive of a toddler, speaking in commanding fashion on a satellite hook-up, for example). But even without entering into discussions of granularity of analysis and combinatorial possibility, it is evident that there is an enormous gap between the few synthetic voices available now, and the half million or so (10*5*11*6*6*7*4) theoretically possible voices listed in Table 2.

Table 2. Theoretically Possible Voices

Parameter	Instantiations	N
Age	infant, toddler, young child, older child, adolescent, young adult, middle-aged adult, mature adult, fit older adult, senescent adult	10
Gender	very male (long vocal tract), male (shorter vocal tract), difficult-to-tell (medium vocal tract), female (short vocal tract), very female (very short vocal tract)	5
Psychological disposition	sleepy-voiced, very calm, calm-and-in-control, alert, questioning, interested, commanding, alarmed, stressed, in distress, elated	11
Degree of formality	familiar, amicable, friendly, stand-offish, formal, distant	6
Size of audience	alone, one person, two persons, small group, large group, huge audience	6
Type of communication	visual - close up, visual - some distance, visual - great distance, visual - teleconferencing, audio - good connection, audio - bad connection, delayed feedback (satellite hook-ups)	7
Communicative context	totally quiet, some background noise, noisy, very noisy	4

Impediments to New Styles and New Voices

We must conclude from this that our current technology provides clearly too few styles of speech and too few voices and voice timbres. The reason behind this deficiency can be found in a central characteristic of TD-synthesis. It will be recalled that this type of synthesis is not much more than a smartly selected, adaptively chained and prosodically modified rendering of pre-recorded speech segments. By definition, any new segment appearing in the synthetic speech chain must initially be placed into the stimulus material, and must be recorded and stored away before it can be used.

It is this encoding requirement that limits the current availability of styles and voices. Every new style and every new voice must be stored away as a full sound database before it can be used, and a 'full sound database' is minimally constituted of all sound transitions of the language (diphones, polyphones, etc.). In French, there are some 2 000 possible diphones, in German there are around 7 500 diphones, if differences between accented/unaccented and long/short variants of vowels are taken into account. This leads to serious storage and workload problems. If a typical French diphone database is 5 Mb, DBs for 'just' 100 styles and 10 000 voices would require (100*10 000*5) 5 million Mb, or 5 000 Gb. For German, storage requirements would double. The work required to generate all these databases in the contemporary fashion is just as gargantuan. Under favourable circumstances, a well-equipped speech synthesis team can generate an entirely new voice or a new style in a few weeks. The processing of the database itself only takes a few minutes, through the use of automatic speech recognition and segmentation tools. Most of the encoding time goes into developing the initial stimulus material, and into training the automatic segmentation device.

And there lies the problem. For many styles and voices, the preparation phase is likely to be much more work than supporters of this approach would like to admit. Consider for example that some speech rate manipulations give totally new sound transitions that must be foreseen as a full co-articulatory series in the stimulus materials (*i.e.*, the transition in question should be furnished in all possible left and right phonological contexts). For example, there are...

- reductions, contractions and agglomerations. In rapidly pronounced French, for example, the sequence 'l'intention *d'*allumer' can be rendered as /*n*alyme/, or 'pe*nd*ant' can be pronounced /pãn^{d}ã/ instead of /pã*nd*ã/ (Duez, this volume). Detailed auditory and spectrographic analyses have shown that transitions involving partially reduced sequences like /n^{d}/ cannot simply be approximated with fully reduced variants (*e.g.*, /n/). In the context of a high-quality synthesis, the human ear can tell the difference (Local, 1994). Consequently, contextually complete series of stimuli must be foreseen for transitions involving /n^{d}/ and similarly reduced sequences.

- systematic non-linguistic sounds produced in association with linguistic activity. For example, the glottal stop can be used systematically to ask for a turn (Local, 1997). Such uses of the glottal stop and other non-linguistic sounds are not generally encoded into contemporary synthesis databases, but must be planned for inclusion in the next generation of high-quality system databases.

- freely occurring variants: '*of the* time' can be pronounced /əvðətajm/, /əvətajm/, /əðətajm/, or /ənətajm/ (Ogden et al., 1999). These variants, of which there are quite a few in informal language, pose particular problems to automatic recognition systems due to the lack of a one-to-one correspondence between the articulation and the graphemic equivalent. Specific measures must be taken to accommodate this variation.

- dialectal variants of the sound inventory. Some dialectal variants of French, for example, systematically distinguish between the initial sound found in '*un* signe' (a sign) and '*in*signe' (badge), while other variants, such as the French spoken by most young Parisians, do not. Since this modifies the sound inventory, it also introduces major modifications into the initial stimulus material.

None of these problems is extraordinarily difficult to solve by itself. The problem is that special case handling must be programmed for many different phonetic contexts, and that such handling can change from style to style and from voice to voice. This brings about the true complexity of the problem, particularly in the context of full, high-quality databases for several hundred styles, several hundred languages, and many thousands of different voice timbres.

Automatic Processing as a Solution

Confronted with these problems, many researchers appear to place their full faith in automatic processing solutions. In many of the world's top laboratories, stimulus material is no longer being carefully prepared for a scripted recording session. Instead, hours of relatively naturally produced

speech is recorded, segmented and analysed with automatic recognition algorithms. The results are down-streamed automatically into massive speech synthesis databases, before being used for speech output. This approach follows the argument that 'if a child can learn speech by automatic extraction of speech features from the surrounding speech material, a well-constructed neural network or hidden markov model should be able to do the same.'

The main problem with this approach is the cross-referencing problem. Natural language studies and psycholinguistic research indicate that in learning speech, humans cross-reference spoken material with semantic references. This takes the form of a complex set of relations between heard sound sequences, spoken sound sequences, structural regularities, semantic and pragmatic contexts, and a whole network of semantic references (see also the subjective dimension of speech described by Caelen-Haumont, this volume). It is this complex network of relations that permits us to identify, analyse, and understand speech signal portions in reference to previously heard material and to the semantic reference itself. Even difficult-to-decode portions of speech, such as speech with dialectal variations, heavily slurred speech, or noise-overlaid signal portions can often be decoded in this fashion (see *e.g.*, Greenberg, 1999).

This network of relationships is not only perceptual in nature. In speech production, we appear to access part of the same network to produce speech that transmits information faultlessly to listeners despite massive reductions in acoustic clarity, phonetic structure and redundancy. Very informal forms of speech, for example, can remain perfectly understandable for initiated listeners, all while showing considerably obscured segmental and prosodic structure. For some strongly informal styles, we do not even know yet how to segment the speech material in systematic fashion, or how to model it prosodically [4]. The enormous network of relations rendering comprehension possible under such trying circumstances takes a human being twenty or more years to build, using the massive parallel processing capacity of the human brain.

Current automatic analysis systems are still far from that sort of processing capacity, or from such a sophisticated level of linguistic knowledge. Only relatively simple relationships can be learned automatically, and automatic recognition systems still derail much too easily, particularly on rapidly-pronounced and informal segments of speech. This in turn retards the creation of databases for the full range of stylistic and vocal variations that we humans are familiar with.

Challenges and Promises

We are thus led to argue (a) that the dominant TD technology is too cumbersome for the task of providing a full range of styles and voices, and (b) that current automatic processing technology is not up to generating automatic databases for many of the styles and voices that would be desirable in a wider synthesis application context. Understandably, these positions may not be very popular in some quarters. They suggest that after a little spurt during which a few more mature adult voices and relatively formal styles will become available with the current technology, speech synthesis research will have to face up to some of the tough speech science problems that were temporarily left behind. The problem of excessive complexity, for example, will have to be solved with the combined tools of a deeper understanding of speech variability and more sophisticated modelling of various levels of speech generation. Advanced spectral synthesis techniques are also likely to be part of this effort, and this is what we turn to next.

Major Challenge One: Advanced Spectral Synthesis Techniques

Reports of my death are greatly exaggerated', said Mark Twain, and similarly, spectral synthesis methods were probably buried well before they were dead. To mention just a few teams who have remained active in this domain throughout the 1990's: Ken Stevens and his colleagues at MIT and John Local at the University of York (UK) have continued their remarkable investigations on formant synthesis (Stevens, 1998; Local, 1994, 1997). Some researchers, such as Prof. Hoffmann's team in Dresden, have put formant synthesisers on ICs. Prof. Vich's team in Prague has developed advanced LPC-based methods, LPC is also the basis of the SRELP algorithm for prosody manipulation, as an

alternative to PSOLA technique, described by Erhard Rank in this volume. Prof. Burileanu's team in Rumania, as well as others, have pursued solutions based on the CELP algorithm. Prof. Kubin's team in Vienna (now Graz), Steve McLaughlin at Edinburgh and Donald Childers/José Principe at the University of Florida have developed synthesis structures based on the Non-linear Oscillator Model. And perhaps most prominent has been the work on harmonics-and-noise modelling (HNM) (Stylianou, 1996; and articles by Bailly, Banga, O'Brien and colleagues in this volume). HNM provides acoustic results that are particularly pleasing, and the key speech transform function, the harmonics+noise representation, is relatively easy to understand and to manipulate [5].

For a simple analysis - re-synthesis cycle, the algorithm proceeds basically as follows (precise implementations vary): narrow-band spectra are obtained at regular intervals in the speech signal, amplitudes and frequencies of the harmonic frequencies are identified, irregular and unaccounted-for frequency (noise) components are identified, time, frequency and amplitude modifications of the stored values are performed as desired, and the modified spectral representations of the harmonic and noise components are inverted into temporal representations and added linearly. When all steps are performed correctly (no mean task), the resulting output is essentially 'transparent', i.e., indistinguishable from normal speech. In the framework of the COST-258 signal generation test array (Bailly, this volume), several such systems have been compared on a simple F0-modification task (www.icp.inpg.fr/cost258/evaluation/server/cost258_coders.html). The results for the HNM system developed by Eduardo Banga of Vigo in Spain are given in sound examples Vigo (a-f). Using this technology, it is possible to perform the same functions as those performed by TD synthesis, at the same or better levels of sound quality.

Crucially, voice and timbre modifications are also under programmer control, which opens the door to the substantial new territory of voice/timbre modifications, and promises to drastically reduce the need for separate DBs for different voices [6]. In addition, the HNM (or similar) spectral transforms can be rendered storage-efficient. Finally, speed penalties that have long disadvantaged spectral techniques with respect to TD techniques, have recently been overcome through the combination of efficient algorithms and the use of faster processor speeds. Advanced HNM algorithms can for example output speech synthesis in real time on computers equipped with 300+ MHz processors.

Major Challenge Two: The Modelling of Style and Voice

But building satisfactory spectral algorithms is only the beginning, and the work required to implement a full range of style or voice modulations with such algorithms is likely to be daunting. Sophisticated voice and timbre models will have to be constructed to enforce 'voice credibility' over voice/timbre modifications. These models will store voice and timbre information abstractly, rather than explicitly as in TD-synthesis, in the form of underlying parameters and inter-parameter constraints.

To handle informal styles of speech in addition to more formal styles, and to handle the full range of dialectal variation in addition to a chosen norm, a set of complex language use, dialectal and sociolinguistic models must be developed. Like the voice/timbre models, the style models will represent their information in abstract, underlying and inter-parameter constraint form. Only when the structural components of such models are known, will it become possible to employ automatic recognition paradigms to look in detail for the features that the model expects [7]. Voice/timbre models as well as language use, dialectal and sociolinguistic models will have to be created with the aid of a great deal of experimentation, and on the basis of much traditional empirical scientific research.

In the long run, complete synthesis systems will have to be driven by empirically-based models that encode the admirable complexity of our human communication apparatus. This will involve clarifying the theoretical status of a great number of parameters that remain unclear or questionable in current models. Concretely, we must learn to predict style-, voice- and dialect-induced variations both at the detailed phonetic and prosodic levels before we can expect our synthesis systems to provide natural-sounding speech in a much larger variety of settings.

But the long-awaited pay-off will surely come. The considerable effort delineated here will gradually begin to let us create virtual speech on a par with the impressive visual virtual worlds that exist already.

While these results are unlikely to be 'just around the corner', they are the logical outcomes of the considerable further research effort delineated here.

A New Research Tool: Speech Synthesis as a Test of Linguistic Modelling

A final development to be touched upon here is the use of speech synthesis as a scientific tool with considerable impact. In fact, speech synthesis is likely to help advance the described research effort more rapidly than traditional tools would. This is because modelling results are much more compelling when they are presented in the form of audible speech than in the form of tabular comparisons or statistical evaluations. In fact, it is possible to envision speech synthesis becoming elevated to the status of an obligatory test for future models of language structure, language use, dialectal variation, sociolinguistic parametrisation, as well as timbre and voice quality. The logic is simple: if our linguistic, sociolinguistic and psycholinguistic theories are solid, it should be possible to demonstrate their contribution to the greater quality of synthesised speech. If the models are 'not so hot', we should be able to hear that as well.

The general availability of such a test should be welcome news. We have long waited for a better means of challenging a language-science model than saying that 'my *p*-values are better than yours' or 'my informant can say what your model doesn't allow'. Starting immediately, a language model can be run through its paces with many different styles, stimulus materials, speech rates, and voices. It can be caused to fail, and it can be tested under rigorous controls. This will permit even external scientific observers to validate the output of our linguistic models. After a century of sometimes wild theoretical speculation and experimentation, linguistic modelling may well take another step towards becoming an externally accountable science, and that despite its enormous complexity. Synthesis can serve to verify analysis.

Conclusion

Current speech synthesis is at the threshold of some vibrant new developments. Over the past ten years, improved prosodic models and concatenative techniques have shown that high-quality speech synthesis *is* possible. As the coming decade pushes current technology to its limits, systematic research on novel signal generation techniques and more sophisticated phonetic and prosodic models will open the doors towards even greater naturalness of synthetic speech appropriate to a much greater variety of uses. Much work on style, voice, language and dialect modelling waits in the wings, but in contrast to the somewhat cerebral rewards of traditional forms of speech science, much of the hard work in speech synthesis is sure to be rewarded by pleasing and quite audible improvements in speech quality.

Footnotes

[1] A diphone extends generally from the middle of one sound to the middle of the next. A polyphone can span larger groups of sounds, e.g., consonant clusters. Other frequent configurations are demi-syllables, tri-phones and "largest possible sound sequences" (Bhaskararao, 1994). Another important configuration is the construction of carrier sentences with "holes" for names and numbers, used in announcements for train and airline departures and arrivals.

[2] LAIPTTS is the speech synthesis system of the author's laboratory (LAIPTTS-F for French, LAIPTTS-D for German).

[3] Interestingly, a speech stretch recreated on the basis of the natural timing measures, but implementing our own melodic model, was auditorily much closer to the original (LAIPTTS g, CD-ROM). This illustrates a number of points to us: First, that the modelling of timing and fundamental frequencies are largely independent of each other, second, that the modelling of timing should probably precede the modelling of F0 as we have argued, and third, that our stochastically derived F0 model is not unrealistic.

[4] Sound example Walker & Local (CD-ROM) illustrates this problem. It is a stretch of informal conversational English between two UK university students, recorded under studio conditions. The transcription of the passage, agreed upon by two native-dialect listeners, is as follows: "*I'm* gonna save that and water my plant with it (1.2 s pause with in-breath), give some to Pip (0.8 s pause), 'cos we were trying, 'cos it says that it shouldn't have treated water." The spectral structure of this passage is very poor, and we submit that

current automatic recognition systems would have a very difficult time decoding this material. Yet the person supervising the recording reports that the two students never once showed any sign of not understanding each other. (Thanks to Gareth Walker and John Local, University of York, UK, for making the recording available.)

[5] A new European project has recently been launched to further research in the area of non-linear speech processing (COST 277).

[6] It is not clear yet if just *any* voice could be generated from a single DB at the requisite quality level. At current levels of research, it appears that at least initially, it may be preferable to create DBs for "families" of voices.

[7] The careful reader will have noticed that we are *not* suggesting that the positive developments of the last decade be simply discarded. Statistical and neural network approaches will remain our main tools for discovering structure and parameter loading coefficients. Diphone, polyphone, etc. databases will remain key storage tools for much of our linguistic knowledge. And automatic segmentation systems will certainly continue to prove their usefulness in large-scale empirical investigations. We *are* saying, however, that TD-synthesis is not up to the challenge of future needs of speech synthesis, and that automatic segmentation techniques need sophisticated theoretical guidance and programming to remain useful for building the next generation of speech synthesis systems.

Acknowledgements

Grateful acknowledgement is made to the Office Fédéral de l'Education (Berne, Switzerland) for supporting this research through its funding in association with Swiss participation in COST 258, and to the University of Lausanne for funding a research leave for the author, hosted in Spring 2000 at the University of York. Thanks are extended to Brigitte Zellner Keller, Erhard Rank, Mark Huckvale and Alex Monaghan for their helpful comments.

References

Bhaskararao, P. (1994). Subphonemic segment inventories for concatenative speech synthesis. In E. Keller (ed.). *Fundamentals in Speech Synthesis and Speech Recognition* (pp. 69-85). Wiley.

Campbell, W.N. (1992a). *Multi-level Timing in Speech.* PhD thesis, University of Sussex.

Campbell, W.N. (1992b). Syllable-based segmental duration. In G. Bailly et al. (Eds.), *Talking Machines: Theories, Models, and Designs* (pp. 211-224). Elsevier Science Publishers.

Campbell, W.N. (1996). CHATR: A high-definition speech resequencing system. *Proceedings 3rd ASA/ASJ Joint Meeting* (pp. 1223-1228). Honolulu, Hawaii.

Greenberg, S. (1999). Speaking in shorthand: A syllable-centric perspective for understanding pronunciation variation. *Speech Communication, 29,* 159-176.

Keller, E. (1997). Simplification of TTS architecture vs. operational quality. *Proceedings of EUROSPEECH '97.* Paper 735. Rhodes, Greece.

Keller, E., & Zellner, B. (1996). A timing model for fast French. *York Papers in Linguistics, 17,* 53-75. University of York. (available at www.unil.ch/imm/docs/LAIP/pdf.files/ Keller-Zellner-96-YorkPprs.pdf).

Keller, E., Zellner Keller, B. & Local, J. (in press). A serial prediction component for speech timing. In W. Sendlmeir (Ed.), *Forum Phoneticum 69: Speech and Signals - Aspects of Speech Synthesis and Automatic Speech Recognition* (pp. 41-50). Frankfurt a. M.: Hector Verlag.

Keller, E., Zellner, B., & Werner, S. (1997). Improvements in prosodic processing for speech synthesis. Proceedings of *Speech Technology in the Public Telephone Network: Where are we Today?* Rhodes, Greece.

Keller, E., Zellner, B., Werner, S., & Blanchoud, N. (1993). The prediction of prosodic timing: Rules for final syllable lenthening in French. *Proceedings ESCA Workshop on Prosody* (pp. 212-215). Lund, Sweden.

Klatt, 1989. Review of text-to-speech conversion for English. *Journal of the Acoustical Society of America, 82,* 737-793.

Klatt, D.H., & Klatt, L.C. (1990). Analysis, synthesis, and perception of voice quality variations among female and male talkers. *Journal of the Acoustical Society of America, 87,* 820-857.

LAIPTTS (a-l). LAIPTTS_a_VersaillesSlow.wav., LAIPTTS_b_VersaillesFast.wav, LAIPTTS_c_VersaillesAcc.wav, LAIPTTS_d_VersaillesHghAcc.wav, LAIPTTS_e_Rhythm_fluent.wav, LAIPTTS_f_Rhythm_disfluent.wav, LAIPTTS_g_BerlinDefault.wav, LAIPTTS_h_BerlinAdjusted.wav, LAIPTTS_i_bonjour.wav..._l_bonjour.wav. Accompanying CD-ROM.

Local, J. (1994). Phonological structure, parametric phonetic interpretation and natural-sounding synthesis. In E. Keller (Ed.), *Fundamentals in Speech Synthesis and Speech Recognition* (pp. 253-270). Wiley.

Local, J. (1997). What some more prosody and better signal quality can do for speech synthesis. *Proceedings of Speech Technology in the Public Telephone Network: Where are we Today?* (pp. 77-84). Rhodes, Greece.

Ogden, R. Local, J. & Carter, P. (1999). Temporal interpretation in ProSynth, a prosodic speech synthesis system. In J.J. Ohala, Y. Hasegawa, M. Ohala, D. Granville, & A.C. Bailey (Eds.), *Proceedings of the XIVth International Congress of Phonetic Sciences, vol. 2* (pp. 1059-1062). University of California, Berkeley, CA.

Riley, M. (1992). Tree-based modelling of segmental durations. In G. Bailly et al., (Ed.). *Talking Machines: Theories, Models, and Designs* (pp. 265 - 273). Elsevier Science Publishers.

Stevens, K.N. (1998). *Acoustic Phonetics.* The MIT Press.

Styger, T., & Keller, E. (1994). Formant synthesis. In E. Keller (Ed.), *Fundamentals in Speech Synthesis and Speech Recognition* (pp. 109-128). Wiley.

Stylianou, Y. (1996). *Harmonic Plus Noise Models for Speech, Combined with Statistical Methods for Speech and Speaker Modification.* PhD. Thesis, École Nationale des Télécommunications, Paris.

van Santen, J.P.H., & Shih, C. (2000). Suprasegmental and segmental timing models in Mandarin Chinese and American English. *JASA, 107,* 1012-1026.

Vigo (a-f). Vigo_a_LesGarsScientDesRondins_neutral.wav, Vigo_b_LesGarsScientDesRondins_question.wav, Vigo_c_LesGarsScientDesRondins_slow.wav, Vigo_d_LesGarsScientDesRondins_surprise.wav, Vigo_e_LesGarsScientDesRondins_incredul.wav, Vigo_f_LesGarsScientDesRondins_itsEvident.wav. Accompanying CD-ROM.

Walker, G., & Local, J. Walker_Local_InformalEnglish.wav. Accompanying CD-ROM.

YorkTalk (a-c). YorkTalk_sudden.wav, YorkTalk_yellow.wav, YorkTalk_c_NonSegm.wav. Accompanying CD-ROM.

Zellner Keller, B., & Keller, E. (in press). The chaotic nature of speech rhythm: Hints for fluency in the language acquisition process. In Ph. Delcloque & V. M. Holland (Eds.). *Integrating Speech Technology in Language Learning.* Swets & Zeitlinger.

Zellner, B. (1996). Structures temporelles et structures prosodiques en français lu. *Revue Française de Linguistique Appliquée: La communication parlée, 1,* 7-23.

Zellner, B. (1997). Fluidité en synthèse de la parole. In E. Keller & B. Zellner (Eds.), *Les défis actuels en synthèse de la parole. Études des Lettres, 3* (pp.47-78). Université de Lausanne.

Zellner, B. (1998a). *Caractérisation et prédiction du débit de parole en français. Une étude de cas.* Thèse de Doctorat. Faculté des Lettres, Université de Lausanne. (Available at www.unil.ch/imm/docs/LAIP/ps.files/ DissertationBZ.ps).

Zellner, B. (1998b). Temporal structures for fast and slow speech rate. *ESCA/COCOSDA Third International Workshop on Speech Synthesis* (pp. 143 - 146). Jenolan Caves, Australia.