



Speech Synthesis in Language Learning: Challenges and Opportunities

Eric Keller and Brigitte Zellner Keller

Laboratoire d'analyse informatique de la parole (LAIP), IMM-Lettres, University of Lausanne, 1015 Lausanne, Switzerland

Eric.Keller@imm.unil.ch, Brigitte.ZellnerKeller@imm.unil.ch

Abstract

This contribution presents an overview of the state of the art of speech synthesis, and its implications for the speech and language sciences. It concentrates on three key arguments. First, it will be shown that speech synthesis quality has advanced to the point where it can be useful for a number of L1 and L2 learning support purposes. Secondly, limits of the dominant time-domain concatenative speech synthesis technology will be demonstrated and explained, and a promising alternative technology will be illustrated. Third, challenges and opportunities arising from these developments for research in speech science and language learning will be delineated^{1,2}.

1. The Use of Current Systems

Many speech synthesis systems have recently 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 (sound examples [SE] 1-3). In this section, we will briefly delineate the main reasons behind these improvements, as well as some of the extended applications that are possible with this technology.

The marked improvements in *signal quality* of good contemporary systems are mainly due to the use of, and improved control over, *concatenative speech technology*, while the *improved coherence* and *greater naturalness* of synthetic speech is primarily a function of our much improved *prosodic modelling*. We shall briefly examine these two developments.

1.1 Concatenative Speech Synthesis and Signal Quality

Early speech synthesis devices (Klatt-Talk, DEC-Talk, or the first generation of InfoVox synthesisers, SE 4-5)

used a technology called "*formant synthesis*" [1-3]. In this approach, voiced speech sounds are created much as one would create a sculpture from stone or wood. For voiced sounds such as vowels or nasals, a complex waveform of harmonic frequencies is created first, and "the parts that are too much", *i.e.* non-formant frequencies, are subsequently suppressed through filtering techniques. For unvoiced or partially voiced sounds, various types of noise are created, or mixed with the voiced signal.

This approach posed several difficulties, the main being that of excessive complexity. Although theoretically capable of producing more or less human-like speech (SE 6-7), these devices must be fed a complex and coherent set of parameters every 2-10 ms. The 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 [4]. But others are apparently language- and speaker-specific, and are more difficult to identify, to implement, and to enforce automatically for internal coherence.

The apparent panacea for these problems has been the generalised transition to an alternative technology called "*time-domain concatenative speech synthesis*" (TD-synthesis for short). In this approach, large databases are collected, and the various constituent speech portions (segments, syllables, words, and phrases) are identified. During the synthesis phase, designated signal portions (diphones, polyphones, etc.³) are retrieved from the database according to phonological selection criteria (a processing step called "unit selection"), they are chained together ("concatenated"), and they are modified for timing and melody ("prosodic modification"). Because such speech portions are basically nothing else than stored segments of actual 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 TD speech synthesis.

¹ While the views developed here may not be uncontroversial in some quarters, they are based on ten years of systematic experimentation with a series of prosodic models and speech synthesis technologies. They also form the basis for extensive portions of a successful speech synthesis system for French and German (LAIP-TTS-F and LAIP-TTS-D). LAIP-TTS-F is available at www.unil.ch/imm/docs/LAIP/LAIP-TTS.html.

² The sound examples for this presentation are available at www.unil.ch/imm/LAIP/LAIP-TTS_challenges.htm. They are marked "SE 1", "SE 2", etc. in the text.

³ 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". [18, 19].

The improvement is remarkable, and we currently use it in our own speech synthesis. At the same time, we do call it a "panacea", because we estimate that TD-synthesis will ultimately prove to be just a transitory technology. This argument will be fully developed in a Section 2, but first we turn to improvements in prosodic modelling.

1.2 Prosodic Modelling and Naturalness of Speech

Prosodic models tend to fall into two categories, strongly linguistic and strongly empirical-statistic ("stochastic") models. For many languages, the initially-developed linguistic models did not furnish satisfactory results, since they were incapable of providing credible predictive timing schemas or the full texture of a melodic line. Our own writings have criticised the exclusive dependence on syntax for the prediction of major and minor phrase boundaries [5], the difficulty of recreating specific Hertz values for the fundamental frequency (melody) line [6], and the exclusive dependence on the notion of "accent" in languages like French where accents are not reliably defined [5].

As a consequence of these inadequacies, stochastic models have moved into dominant position among high-quality speech synthesis devices. These models generally implement an arborescence of predictive parameters, and derive statistical coefficients from extensive database material. The predictive parameters themselves (such as a position in the syllable, the word and the phrase, sounds making up the containing syllable, preceding and succeeding sounds, the lexical status of the containing word, etc.) do not change a great deal from language to language, or from project to project [7, 8].

Differences between stochastic models concern the priority of either timing or melody modelling, the precise statistical approach used for modelling (e.g., artificial neural network, classification and regression tree, sum-of-products model, general linear model [9-11]), and the logic underlying the arborescence.

In our own writings, we have suggested that the modelling of the precise melody line could well depend on the knowledge of correct timing values, and that the modelling of timing should thus precede the modelling of fundamental frequency [5, 6]. The question of whether timing or fundamental frequency modelling comes first has to our knowledge never been systematically investigated.

Also, we have shown that psycholinguistically driven dependency trees oriented towards actual human speech behaviour give better predictive results than dependency trees derived from purely syntactic principles [5]. Moreover, recent work has shown that this latter result is not a peculiarity of our work on French, because largely similar results have also been obtained with German [12].

A typical contemporary high-quality system thus tends to be a TD-synthesis system incorporating two stochastic models, one for timing and one for melody. Grapheme-to-phoneme and phonological processing tends to be rule-driven. Our own system for French (LAIPTTS-F) is not unusual in that sense. It incorporates 520 grapheme-phoneme rules, a set of phonotactic rules implementing a small proximal grammar (liaison, enchaînement, homograph processing, etc.), a 7000+-item dictionary of irregular pronunciations of common words, and a prosodic module incorporating a statistical model of some 11'000 hand-measured phoneme segments in 277 sentences of read speech that incorporate 98.5% of French transitions [13]. In the next section, we will demonstrate some of the uses that such a system can be put to in various language learning contexts.

1.3 Uses for High Quality Speech Synthesis

We estimate that many speech synthesis systems can now be used in a series of concrete applications:

- Assisting the *language teacher* in certain language learning exercises, including as an aid to reading. Speech synthesis allows repetition at will, as well as the presentation of exercises specifically adapted to the needs of the student, plus the creation of sound examples that could *not* be produced by a human being (e.g., speech with intonation, but no rhythm).
- Assisting *researchers in linguistics or psychology* in producing speech stimulus material in specific and controlled ways, in order to test theoretical hypotheses. In this sense, speech synthesis is becoming an interesting experimental tool that favours the development of objective methods (i.e., the use of instrumental, rather than impressionistic approaches). This encourages reproducibility of experimental results between laboratories.
- Simulating a "serious and responsible speaker" in various *virtual environments* (e.g., friendly helper's voice for the visually handicapped, a news reader in a virtual radio station, a speaker of a extinct and recreated language, or a salesman in a virtual store, etc.).

Let us evoke some of these applications in greater detail.

1.4 Language Teaching

A good and flexible model of a given target language can be useful for the training of *prosodic and articulatory competence*. Speech synthesisers can slow down stretches of spoken language at will, which eases familiarisation and articulatory training with novel sound sequences. Learners can begin with speech sequences that are produced slowly, and increase the speed as their facility improves (SE 9-11).

Advanced learners can experiment with the accelerated reproduction speeds also used by the visually handicapped for scanning texts (SE 12-13). English-speaking learners of French, for example, need considerable training to integrate French rhythm, which diverges from the stress patterns characteristic of English. (The inverse is also true, of course). New examples can be generated at will with synthesis, even in the absence of a native speaker.

Listening comprehension is another obvious L2 application area. Many synthesis systems can be stopped in mid-stream, backed up one or several sentences, and can repeat what was just read out, possibly more slowly. Here a speech synthesis system acts as an "*indefatigable substitute native speaker*".

1.5 Training in Reading

A high-quality speech synthesis can also be used for *illiteracy 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 illiterates. Teaching materials can be combined with attractive, game-like interfaces to reinforce the favourable preconditions for such learning. Endowed by a fully-adapted interface, a speech synthesiser could be used as an interactive assistant, required for reinforcing the learning of correspondences between the written and the spoken language.

1.6 Linguistic and Psycholinguistic Experimentation

Speech synthesis can become a useful tool for linguistic and psycholinguistic experimentation, since it permits to incorporate knowledge from selected and diverse levels (phonetic, phonological, prosodic, lexical, etc.), and to verify the relevance of each as they interact with each other. Although the use of speech synthesis in this context is still in its infancy, it is already now possible to perform a number of manipulations with current speech synthesis systems to isolate and experiment with rhythm and pitch patterns, the placement of major and minor phrase boundaries, and typical phonological patterns in a language (SE 14-19). Humans cannot demonstrate these functions separately, since their active control of phonation does not permit it. Each such experiment can isolate an important aspect of prosodic structure, and in so doing can illustrate its contribution to the overall acoustic effect.

1.7 Computer Tool

In recent years, dictionaries, grammars (as correctors), translation systems, and search engines responding to full sentences have found their place on our computers. Speech synthesisers are likely to follow next. When the language competence of a system begins to outstrip that of some of the better second language users, such systems become useful new adjunct tools.

2. The Limits of Current Systems

If current high-quality systems can be put to excellent use in a variety of established and new applications, they are also subject to considerable limitations. In this section, we wish to illustrate some of these limits, explain their origins, and explore some alternatives.

2.1 Not Every Speaker Follows the Same Model

In section 1.2, we said that our prosodic model was based on detailed measurements of a native speaker of French, reading a long series of French sentences. Furthermore, we showed that speech driven by this model is credible and can be useful for a variety of purposes.

However, the model we established for our speaker is evidently not the only possible timing model. Sound example 20, for example, is a short portion taken from a French TV newscast of January 1998, and sound example 21 illustrates the reading of the same text with our speech synthesis system. The prosodic structures of the two examples are very different. Analysis of the example showed, on the one hand, that the two renderings differ primarily with respect to timing, and on the other, that the speaker's temporal structure cannot be easily derived from our timing model⁴. In order to produce a timing model for this speaker, it seems likely that a large portion of the empirical study underlying the original model would have to be redone from scratch (*i.e.*, another 10'000 segments to measure, and another statistical model to build).

2.2 How Many Styles?

TV speaker presentation style is only one of many styles that current speech systems do not necessarily integrate. At the present time, it is impossible to say exactly how many styles are "still missing", but it must be many. Table 1 shows an initial, rough calculation.

The total derived from this list is 180 ($4 \times 5 \times 3 \times 3$) theoretically possible styles. It is true that some styles could probably be modeled as deviations from other styles, and that some theoretical combinations are impossible or unlikely (a spelled, commanding presentation of questions, for example). This list is presented primarily for argument's sake, and not as a solid basis for further research.

⁴ 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 (SE 22). To us, this illustrates a number of points: 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 fundamental frequency as we have been arguing in our writings, and third, that our stochastically derived fundamental frequency model is not unrealistic.

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

But even with these caveats, it remains obvious that only few of all possible styles are supported by current speech synthesis systems. In our own system, for example, we can only deal with reading style, and only one dialect is approximated. We have implemented initial versions for the four speech rates, and for the three types of reading material listed here. Our system is thus capable of a total of just 12 out of a theoretical 100+ styles. More styles would clearly be desirable.

2.3 Truly Expressive Speech is Still in the Future

To illustrate some further limits of current speech synthesis, we analysed a small speech segment from an animated film. In a 1966 *Peanuts* television special [14], Lucy asks Linus to roll an enormous pumpkin into the house. Sweating at the brow, Linus looks on in horror as Lucy seizes an enormous kitchen knife, plunges it into the pumpkin, and extracts the pumpkin seeds. Finally, Linus cries out in great anguish: "Oh...you didn't tell me you were gonna kill it!" (SE 23). Translated into French and pronounced by our speech synthesiser, that gives a most disappointing "Ô ô ô - tu ne m'avais pas dit que tu allais la tuer!" (SE 24).

Emotional and expressive speech constitutes another evident gap in contemporary speech synthesis systems, despite a considerable theoretical effort directed at the question [15]. The lack of general availability of emotional variables prevents current systems from being put to use in animation, automatic dubbing, virtual theatre, etc. Technically, the lack of expressive voices is related to limits on the availability of voices in general. Most current systems provide just a single adult voice. Some offer two or three. But before explaining the origin of these limitations, we shall raise the question of how many voices would be theoretically desirable.

2.4 How Many Voices?

In Table 2, we have made a list of factors that are known to, or can conceivably influence, voice quality. Again, there is no guarantee that this list is complete, or that all theoretically possible combinations are also possible ones (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 grain 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 ($10 \times 5 \times 11 \times 6 \times 6 \times 7 \times 4$) theoretically possible voices that are 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

2.5 Impediments to New Styles and New Voices

So there are clearly too few styles of speech and too few supported voices and voice timbres. The reasons behind this deficiency are to be found in a central characteristic of TD-synthesis.

It will be recalled that this type of synthesis is really 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.

This requirement entails that every new style or every new voice must be stored away *as a full sound database* before it can be used, and that a "full sound database" is minimally constituted of all sound transitions of the language (diphones, polyphones, etc.). In French, there are some 2000 possible diphones, in German there are around 7500 diphones, if differences between accented/unaccented and long/short variants of vowels are taken into account.

This ultimately leads to serious storage and workload problems. If a typical French diphone database is 5 Mb, DB's for "just" 100 styles and 10'000 voices would require (100*10000*5) 5 million Mb, or 5'000 Gb. For German, these storage requirements would double.

The work required to generate all these databases in 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 just a few weeks. The generation 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 speed 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 /nalyme/, or "pendant" can be pronounced /pã^dã/ instead of /pãdã/ [16]. Detailed auditory and spectrographic analyses have shown that transitions involving partially reduced transitions 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 [17]. 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 [18]. 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 /vð*tajm/, /v*tajm/, /ð*tajm/, or /n*tajm/ [19]. 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 "*insigne*" (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. It is the fact that special case handling must be programmed for many such problems, and that such special case handling changes from style to style or voice to voice, which 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.

2.6 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 in preparation of a directed recording session. Instead, hours of relatively naturally produced speech is being recorded under studio conditions, and is 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*. Current psycholinguistic and perceptual research indicates 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. 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 ⁵ (see e.g., [22]).

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 is marked by a specific style and that is appropriate to the specific speech context. This enormous network of relations takes a human being twenty or more years to build, using the most phenomenal parallel processing system in existence today ⁶.

Current automatic analysis systems are still very far from that sort of processing capacity, or such a sophisticated level of linguistic knowledge. Consequently, only relatively simple relationships can be learned automatically, and automatic recognition systems still derail much too easily. This in turn retards the creation of databases for the full range of stylistic and vocal variations that we humans are familiar with. For some informal styles, we do not even know yet how to store the speech material in databases, or how to model it prosodically.

2.7 Towards an Alternative

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 automatic processing technology, very much in vogue today, 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 are not 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 return to spectrally-based technologies. This in turn

⁵ Sound example x 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 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.)

⁶ It is commonplace to admire the processing capacities of the human brain. But here is one particularly pregnant way of stating it: There are about the same number of inter-neural connections in our brain as there are leaves on the ten billion trees in the Amazon basin. [27].

means that we shall have to face up to some of the tough speech science problems that we left temporarily 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.

2.8 Advanced Spectral 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 [4, 17-19]. Some researchers, such as Prof. Hoffmann's team in Dresden, have put formant synthesisers on chips. Prof. Vich's team in Prague has developed advanced LPC-based methods. Prof. Kubin's team in Vienna have developed synthesis structures based on the Non-linear Oscillator Model, and Prof. Burileanu's team in Rumania, as well as others, have pursued solutions based on the CELP algorithm. And perhaps most prominent has been the work on harmonics-and-noise modelling (HNM) [23-26]. HNM modelling has been chosen by AT&T as the basis for their "next-generation speech synthesis", the acoustic results are particularly pleasing, and the key speech transform function, the harmonics+noise representation, is relatively easy to understand and to manipulate.

For a simple analysis-re-synthesis cycle, the algorithm proceeds in the following steps:

- Narrow-band spectra are obtained at regular intervals in the speech signal (once every 2-10 ms).
- Amplitudes and frequencies of the fundamental and harmonic frequencies are identified up to about 8 kHz.
- Irregular and unaccounted-for frequency components are identified and stored separately.
- Time, frequency and amplitude modifications of the stored values are performed as desired.
- 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 totally "transparent", i.e., indistinguishable from normal speech. In the framework of the COST-258 signal generation test array, 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 25-30.

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 largely suppresses the need for separate DB's for different voices⁷. In addition, the HNM (or similar) spectral transforms are storage-efficient. The problem of excessive duplication of databases is thus likely to find its natural absorption, as the field evolves towards a generalised use of advanced spectral techniques. 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.

3. Challenges and Promises

A picture of considerable vibrancy has been sketched here of the current status of speech synthesis. It is time to present our own personal view of these trends, and to show challenges posed and promises made by speech synthesis for the speech sciences in general.

3.1 The Challenges: Create the Models

- The search for even greater naturalness, and the need for more speech styles, voices and timbres for a many more languages will be the driving forces behind speech synthesis development during the coming years.
- It will not be possible to meet these challenges with current TD-synthesis techniques. Advanced spectral techniques will be required to provide the full range of voice and timbre modifications. Once these techniques are in general use, *sophisticated voice and timbre models* will have to be constructed to enforce "voice credibility" in voice/timbre modifications. These models will store voice and timbre information *abstractly*, rather than *explicitly* as in TD-synthesis, in the form of underlying parametric contributors and inter-parameter constraints.
- It will not be possible to satisfy the full range of speech style requirements with automatic processing techniques. To handle informal styles of speech in addition to more formal styles, and to handle the full range of dialectal variation in

⁷ 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 DB's for "families" of voices.

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. Once developed, *automatic recognition paradigms can be programmed* to look in detail for the features that the model expects⁸.

This leads to the key message to this language- and speech-oriented conference audience: Contrary to what has been touted extensively, the voice/timbre models as well as the language use, dialectal and sociolinguistic models *are unlikely to "emerge spontaneously"* from the computers of our engineering colleagues. These 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.

3.2 Promise 1: Opportunities for Linguistic Science

If we, the members of the language and speech science community, choose to work with the computational and statistical tools that are commonplace in engineering, we can contribute intelligently to the advancement of this domain. In the authors' minds, the time has come to begin driving complete synthesis systems with empirically-based models that encode the admirable complexity of our human communication tools. This will also contribute to clarifying the theoretical status of a great number of parameters that remain unclear or questionable in current models.

3.3 Promise 2: Towards Greater Scientific Accountability

In addition, we suspect that speech synthesis will during the coming years become elevated to the status of an *acid test* for our models of language structure, language use, dialectal variation, sociolinguistic parametrisation, as well as timbre and voice quality. If the model is good, it will sound good in prolonged synthesised sequences, even to non-specialist ears. If the model is not so hot, we will hear that as well.

And that is just the way it should be. We have long waited for a better means of challenging a model than saying that "my *p*-values are better than yours" or "my

⁸ The careful reader will have noticed that we are *not* suggesting that the positive developments of the last decade be simply put into the trash. 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.

informant can say what your model doesn't allow". Starting immediately, we can run a language model through its paces with many different styles, stimulus materials, speech rates, and voices. We can cause it to fail, and test it under rigorous controls.

This will permit even general scientific observers to validate the output of our linguistic models. After a century of theoretical speculation and experimentation, linguistic modelling can take another step towards becoming an externally accountable science, despite its enormous complexity.

4. Acknowledgements

Grateful acknowledgement is made to the Office Fédéral de l'Education (Switzerland) for supporting this research through its funding of Swiss participation in COST 258, and to the University of Lausanne and the État de Vaud for funding research leaves for the two authors, hosted in Spring 2000 at the University of York.

5. References

- [1] Klatt, 1989. Review of text-to-speech conversion for English. *Journal of the Acoustical Society of America*, 82, 737-793.
- [2] 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.
- [3] Styger, T., & Keller, E. (1994). Formant Synthesis. In E. Keller (ed.). *Fundamentals in Speech Synthesis and Speech Recognition* (pp. 109-128). Wiley.
- [4] Stevens, K.N. (1998). *Acoustic Phonetics*. The MIT Press.
- [5] Zellner, B. (1998). *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).
- [6] 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. September 1997.
- [7] Keller, E., Zellner-Keller, B., & Local, J. (in press). A serial prediction component for speech timing. In W. Sendlmeier (ed.) *Festschrift Wolfgang Hess*.
- [8] 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.
- [9] Campbell, W.N. (1992). Syllable-based segmental duration. In G. Bailly, & al (Eds.), *Talking Machines. Theories, Models, and Designs* (pp. 211-224). Elsevier Science Publishers.
- [10] 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.
- [11] Keller, E., & Zellner, B. (1996). A timing model for fast French. *York Papers in Linguistics*, 17, University of York. 53-75. (available at www.unil.ch/imm/docs/LAIP/pdf.files/Keller-Zellner-96-YorkPprs.pdf).
- [12] Siebenhaar-Rölli, B. (2000). *The timing model for the Swiss High German LAIPTTS*. COST 258 Meeting, Stockholm, May 2000. (available at www.unil.ch/imm/docs/LAIP/COST_258/Meetings/seventh_stockholm/TiminginSwissGerman.pdf).
- [13] Keller, E. (1997). Simplification of TTS architecture vs. operational quality. *Proceedings of EUROSPEECH '97*. Paper 735. Rhodes, Greece. September 1997.
- [14] Commercially available as "Peanuts Classic: It's the Great Pumpkin, Charlie Brown", Paramount Video.
- [15] See for example ISCA workshop on speech and emotion, 5-7 Sept, 2000. www.qub.ac.uk/en/isca/index.htm.
- [16] Duez, D. (in press). Reduction and assimilatory processes in conversational French speech: Implications for speech synthesis. Final Report, COST 258.
- [17] 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.
- [18] 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?* Rhodes, Greece. September 1997.
- [19] Ogden, R. Local, J. & Carter, P. (1999). Temporal interpretation in ProSynth, a prosodic speech synthesis system. *Proceedings of the XIVth International Congress of Phonetic Sciences* (Ohala, J.J., Hasegawa, Y., Ohala, M., Granville, D., and Bailey, A.C. eds.), 2, pp. 1059-1062. University of California, Berkeley, CA.
- [20] 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.
- [21] Campbell, W. N. (1996). CHATR: A high-definition speech resequencing system. *Proc. 3rd ASA/ASJ Joint Meeting*, 1223-1228. Honolulu, Hawaii.
- [22] Greenberg, S. (1999). Speaking in shorthand: A syllable-centric perspective for understanding pronunciation variation. *Speech Communication*, 29, 159-176.
- [23] 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.
- [24] Bailly, G. (in press). A parametric harmonic+noise model. *Final Report*, COST 258.
- [25] Banga, E.R., Fernández-Salgado, X., & García-Mateo, C. (in press). Concatenative text-to-speech synthesis based on sinusoidal modelling. *Final Report*, COST 258.
- [26] O'Brien D., & Monaghan A.I.C. (in press). Shape invariant pitch and time-scale modification of speech based on a harmonic model. To appear in COST 258 Final report.
- [27] Snellgrove, B. (1996). *The Unseen Self*. Saffron Waldon, Essex: The C.W. Daniel Co.