New Uses for Speech Synthesis

by Eric Keller and Brigitte Zellner Keller
Laboratoire d'analyse informatique de la parole (LAIP)
Faculté des Lettres, Université de Lausanne
1015 LAUSANNE, Switzerland
Eric.Keller@imm.unil.ch and Brigitte.ZellnerKeller@imm.unil.ch

Preface

A team of phoneticians, linguists, computer scientists and mathematicians at our laboratory have over the past few years developed large portions of a new speech synthesis system for French. Its distinguishing aspect is a greatly improved, natural-sounding quality. This system is freely downloadable (for non-commercial purposes) from www.unil.ch/imm/docs/LAIP/LAIPTTS.html, and it works in conjunction with the freely available Mbrola sound output system. Much of this presentation can also be obtained, in English and French, together with some illustrative sound examples, at the following web page: www.unil.ch/imm/docs/LAIP/LAIPTTS_sim.htm.

Speech Synthesis as part of Virtual Reality

For most of us, the term speech synthesis evokes memories of mechanical, monotonous or repetitive voices, while virtual reality makes us think of amazingly realistic film scenes. But in fact, speech synthesis has been much improved in the last ten years, and it is now firmly part of virtual reality. And with that, researchers in the humanities can now use speech synthesis in novel ways.

We estimate that many speech synthesis systems are now ready, among other things, for the following new tasks:

- 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.).

All these uses can be understood as a type of virtual reality, since we *simulate* a human speaker, in whole or in part, for a specific application.

The Limits of Current Speech Synthesis

At the same time, the time is clearly not yet ripe for some other potential uses of speech synthesis:

- Synthetic voices are still *insufficiently expressive*. We cannot yet simulate human emotion, such as joy, anger or sadness satisfactorily. This means that our artificial voices do not yet have the full, extended "vocal palette" of a human speaker required for "virtual theatre" or animated film applications.
- The synthetic *voices themselves* are still very limited. In some European languages, we may have a few male and a few female adult voices for a given language. But where are the children's voices, where are the adolescents, or where are the senior voices? With today's concatenative technology, the creation of a new voice represents a major effort, even for large and well-financed research teams.
- There is by and large still very little *dialectal synthesis*. Also, there is no synthesis of informal language, and no simulation yet of social class variants.

To the members of our laboratory, these are the true challenges of the future. In the following sections, we illustrate in some more detail what good speech synthesis can do at the present, and what we would like it to do in the future.

Speech Synthesis for Language Teaching

The learning of a second language involves a great variety of skills where speech synthesis can be used meaningfully. A good model of the language is particularly useful in 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.

Advanced learners, on the other hand, may wish to experiment with accelerated reproduction speeds. These are commonly used by the visually handicapped for scanning voluminous texts. English-speaking learners of French, for example, need considerable training to integrate French rhythm, which diverges considerably from the stress pattens characteristic of English. New examples can be generated at will with synthesis.

Another obvious area of application concerns *listening comprehension*. It is true that today's learners of major foreign languages can easily connect with Internet-based radio or TV from the desired language area. But at the same time, such transmissions can rarely be halted or slowed down at will. Many synthesis systems can be stopped in mid-stream, can back up one or several sentences, and can repeat what was just read out, possibly much more slowly. (Our system does not have this capacity yet, but will so soon.)

Which leads to yet another use of speech synthesis systems, that of a *"substitute native speaker"*. Dictionaries and grammars (as correctors) increasingly show up on our computers. Why not add a speech synthesiser? When the language competence of the system begins to outstrip that of some of the better second language users, such systems become useful new adjunct tools. In fact, by their ability to produce natural-sounding speech from almost any text, they may soon become as indispensible on one's personal computer as the latest electronic dictionary.

Speech Synthesis for Training in Reading

A high-quality speech synthesis can also be at the basis of tomorrow's tools to combat *illiteracy*. We know that in our developed societies, illiteracy has stigmatizing status. The attraction of having such a tool on a computer is that the computer is precisely *not* a human, and is thus likely to be perceived as non-judgemental and neutral by illiterates. Teaching materials could be combined with attractive, game-like interfaces to reinforce the favorable preconditions for such learning. Endowed by a fully-adapted interface, a speech synthesiser could be used as an indefatigable and interactive repeater and assistant, required for reinforcing the learning of correspondances between the written and the spoken language.

Speech Synthesis for Linguistic and Psycholinguistic Experimentation

Knowledge about language and speech functions has considerably increased during the last twenty years. The overall picture of human linguistic functions is of such complexity that researchers capable of integrating all of this knowledge have become rare. In this sense, 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. Humans cannot demonstrate these functions separately, since their active control of phonation does not permit it. Consider, for example, the following experiments:

Experiment	***How to do it with our synthesiser***
Demonstrate the effects of greater and lesser degree of monotony (fundamental frequency variation)	Adjust "melody" in Control | Parameters.

Demonstrate the effects of greater and lesser speech rate acceleration	Adjust "linear speed" in Control | Parameters
Demonstrate the effects of no fundamental frequency variation at all	• Select "Save Prosodics" in Control | Parameters • Have synthesiser read a passage aloud • Open the resulting *.pho file with Excel • Change all the fundamental frequency values to the same frequency (e.g., 100 Hz) • Have Mbrola read the changed *.pho file
Demonstrate the effects of no rate variation at all	• Select "Save Prosodics" in Control | Parameters • Have synthesiser read a passage aloud • Open the resulting *.pho file with Excel • Change all the timing values to the same duration (e.g., 100 ms per segment) • Have Mbrola read the changed *.pho file
Demonstrate the effects of the absence of word grouping in an utterance	• Open a text • Place commas after each word • Save the text • Have the synthesiser read the text and note the absence of the word grouping effect
Demonstrate the effects of the rhythmic structure of some frequent lexical words	• Create a list of 20 disyllabic words (e.g., armchair, language, Europe, avoid, excel, etc.) • Save the text • Have the synthesiser read out the entire list and listen for the rhythm

Each of these experiments isolates an important aspect of prosodic structure, and in so doing illustrates its contribution to the overall acoustic effect.

Historic Reconstruction

As part of virtual reality, speech synthesis is also likely to become part of another interesting new trend, that of virtual historic reconstruction.

In quite a number of places in Europe, combined knowledge from manuscripts and excavations is being used to reconstruct in precise detail the state of historic sites. At York (UK), for example, a Viking village has been recreated in minute detail on the basis of accumulated evidence and at the place of the original site, *i.e.*, underneath a large shopping centre complex (http://www.jorvik-viking-centre.co.uk/). As visitors pass through the village in electric carts, real-size models of village inhabitants are heard to converse in a form of Old English that was recreated to be as similar as possible to the assumed form of the informal speech of the time. However, when listening carefully to the voices recreated at York, it appears that at least some of the modern speakers used to record the utterances were only partially confortable with the language, since the fluency typical of informal, conversational speech is often lacking.

This is not an isolated case. A careful examination of examples of classical Latin (first century B.C.), recited by a known authority of Latin pronunciation, identified multiple pronunciation errors. This is easily understood: since Latin grapheme-to-phoneme rules, for any given period, differ considerably from those of any modern European language, and since there is no language community that uses classical Latin pronunciation on an every-day basis, it is difficult for a modern speaker to become fully and solidly conversant with its pronunciation. The suspicion lies near that a well-tuned speech synthesis system might soon do better than modern "non-native" speakers of Old English or classical Latin, and provide totally fluent and natural-sounding renditions of "informal Old English" and "Cicero's Latin". This is being tested with respect to recited classical Latin by a doctoral student writing his thesis in association with our laboratory. In preparation for a recreation of Latin synthetic speech, Latin and secondary sources concerning the phoneme inventory and pronunciation of classical Latin are being examined.

Surprising phonetic detail can be inferred from contemporary comments. For example, Cicero complains that Ennius always wrote "Phyrrus" as "Burrus" ("Burrum semper Ennius, numquam Pyrrhum", oratio 160). Quintus Ennius (239-169 B.C), who was later to become an influential Latin epic poet, was a native of Calabria and learned Toscan and Greek in his youth. He grew up at a time when the region was just coming under Roman dominance. This supports the notion that Roman plosives were more strongly aspirated, or less voiced, than corresponding Toscan plosives (*i.e.*, characterised by longer VOTs). At the same time period, Greek loan words that were spelled with initial *phi* in Greek were reliably transliterated into Latin with "ph-". From indices such as these it becomes possible to recreate a "hypothetical VOT line" for the unvoiced plosives of the time, with the longest VOTs associated with Greek, classical Latin VOTs taking an intermediate position, and Toscan VOTs occupying the least aspirated, or possibly even a negative VOT pole.

In an interesting sidelight on the difficulties generating an extinct language synthetically, extinct languages may contain diphones that are difficult or impossible to find in modern diphone bases. For examples, it can be concluded from the frequent omission of final "m" on tombstones as early as 259 B.C. that [-Vm V-] sequences were fully nasalised during most of the Roman period. In fact, Quintilian (1st century AD) comments directly on this by saying (Quint. inst. 9, 4, 40) "Whenever this same letter (*i.e.*, 'm') ends a word and enters into contact with a vowel starting the next word, in a manner that a transition becomes possible, the final 'm' is barely pronounced, all while being maintained in the orthography, for example in *multum ille* and *quantum erat*, so that it gives nearly the impression of a sound of some new letter. In fact, it is not suppressed, but rendered indistinct, and only represents a type of sign that keeps two vowels from being confounded." Assuming one wished to recreate Latin appropriate to Cicero's time, one would probably have to use moderately aspirated or unaspirated plosives combined with nasal vowels. This is not easy to do with currently available diphone databases. French DBs do contain most nasal vowels, as well as unaspirated plosives, but they do not contain the nasal [u]. This means that the reconstitution of Latin may well require building an entirely new diphone database, or signal manipulations within an existing database to create missing diphones.

Synthetic reconstitution of historic languages will also serve to point out some *limits* of our knowledge of extinct languages. For example, what was the prosody of classical Latin? Since members of past civilisations were generally of shorter stature than contemporary humans, does that mean that their average fundamental frequency was higher? There is some indication that this may well be the case, since music written for 15th-16th century male singers often reaches into the very top of the contemporary tenor range. But even if we are led, by skeletal measures and serious correlational analyses, to assume an average fundamental frequency shift by a half an octave or more, we will probably remain in considerable doubt about the *melody contours* that would have to be postulated. In that sense, the synthetic recreation of extinct languages will serve to point out the numerous remaining weaknesses in our current understanding of such languages.

Synthesis of extinct languages is likely to be bound into attempts to recreate virtual historic sites. For example at the University of Caen (France), a multidisciplinary research team has launched a project on the "Virtual reconstitution of Antique Rome" (see http://www.unicaen.fr/rome/). Issuing from the formidable early-20th century effort of Paul Bigot's 70-m^2 3D plaster cast of ancient Rome, the French team is in the process of digitizing and updating the entire cast, in order to provide the backdrop for a "3D virtual Rome" evolving over the centuries. In time, it is hoped that this virtual scene will become a photorealistic stage for recreating the scenes of the classical Roman speeches. We in Lausanne hope to see our project on Latin synthesis contribute to the capacity for producing classical Latin oratory speech with as much realism as possible.

Where are we headed?

Despite all the justified enthusiasm about the improved capacity for increasingly pleasant-sounding synthesised speech, current capacities are still limited. Under good circumstances, we have a credible capacity for a relatively formal reading style. Practically unknown today are systems that give truly expressive speech. Expression of surprise, anxiousness, excitation or disappointment are very difficult to impossible to generate with present-day concatenative synthesis technology.

Quite a few research teams are working on these problems. Many such laboratories are part of the European COST 258 Project, see http://www.unil.ch/IMM/docs/LAIP/COST_258/cost258.htm. It is likely that in a few years' time, further

steps will be taken towards greater realism of artificial voices. With the impressive results of Harmonics and Noise Modelling (HNM) of speech, for example, the technology for building such a capacity is already in place. As HNM systems mature and become available for applications as described here, speech synthesis can be used even better for the purpose of understanding and assisting human communication in multiple novel fashions.

Acknowledgement

Grateful acknowledgement to Olivier Bianchi, University of Lausanne, for detailed information on classical Latin pronunciation.