Speech synthesis development and phonetic research - a personal introduction

Rolf Carlson and Björn Granström

Generations

We are now in the process of creating, educating and indoctrinating a third generation of speech synthesis researchers. We are also in the process of developing a third generation of synthesis systems. These generations go hand in hand and some unique researchers span all these generations as can be seen in this issue. The first generation made a breakthrough by creating systems that produced human like sounds. The output created surprise, applause and enthusiasm among the general public. However, the development work had a deeper meaning than the art of a magician. The research was heading for a deeper understanding of articulation and perception and a description of the speech code. Thus, the driving force was not primarily to make the machine talk but to study phonetics in a broader sense. However, in grant applications the practical importance was emphasised and a bright future was painted. What happened with the future?

The next generation or the next phase introduced computers and the art had to be squeezed into mathematical formalisms. Computer programs were developed to control the analog synthesizers and several strategies were invented. The work continued and success seemed to be just around the corner, but not quite. The programs became too complex to understand and the introduction of the digital simulation of the terminal analog introduced new possibilities and restrictions on the total system. New methods for rule implementation were developed, influenced by new trends in linguistics. The systems were turned into rule based systems. Where were the rules to be found? What had the phoneticians been doing all this time? The well studied details did not fit into the general model that should handle all possible combinations at all levels. We realized that we understood a lot of some things and very little about other things. Unfortunately the latter dominates the perceived general quality of speech synthesis. Even if the first systems were a blessing for some people that urgently needed a synthetic voice, the general reception started to be more reserved or in some cases even negative. The days were over when people were impressed, positive and courteous when they were exposed to speech synthesis.

We are now in the third phase, when the promises should be fulfilled and a breakthrough should be made. In this issue of Journal of Phonetics we have gathered a group of researchers representing the view that phonetics and speech synthesis have, or should have, a lot in common. As already mentioned in the introduction the selection has been made to penetrate the subject from different angles. We must conclude that the selection of authors is mostly done from the first two generations and the message should be clear: make use of the gathered knowledge! The review by Klatt, in JASA 1987, is a landmark for all of us working in speech synthesis. We have not wished to repeat his historical approach in general, but in the paper by Karlsson some specific historical experiments with synthesis of the female voice are described.

Old knowledge and new methods?

Where does the researcher find ideas and new knowledge? Each laboratory or teaching place has its own traditions, methods and assumptions. A "new" idea is mostly the result of these sources and sometimes a glance over the fence into the

neighbour's yard. If we study this pattern in more detail, we find that a new idea is more like a ripe fruit to be harvested than a new path into the unknown. Small steps push us in a certain direction and each step is a consequence of the past. This is not a negative trend as such, but we do not always realise what is happening and how everything fits together. We can see many examples of work that was done because it was simply the correct time to do it: a critical mass had been reached in the research community.

By forgetting the past and by sidestepping other people's work, we reduce the scientific quality of our own work and we limit our possibilities to understand when we actually are breaking new ground. The speech community can not in the long run afford to rediscover old facts and methods.

We find an attenuation in the traditionally strong line of phonetic/synthesis research so well mastered by the Haskins group: the detailed study of phonetic cues. The perceptual work that mapped the acoustic space into phonetic categories depending on context in its widest sense or expectation is currently a small vein. The broad views and the complex systems are more in style. We find this to be very alarming since, for example, the details of coarticulation rules form an important foundation for a synthesis system.

On the other hand, new methods from other research areas are coming into the synthesis field. For example, nonlinear phonology is slowly being recognized by synthesis researchers. The work by Hertz in this issue is an example of this trend. Parallel processing, artificial neural networks and statistical models that are common in speech recognition are finding new applications in our area. We have already seen examples of grapheme-to-phoneme conversion based on neural networks and syntax analysis based on statistical models.

Use of speech and language corpora

Many of the new methods need much training data on acoustical and linguistic levels. Huge corpora of speech and text are currently being collected and labelled. The TIMIT data base is an early example in this direction. Recorded speech is labelled according to many different philosophies, however. If the assigned labels are used without due consideration, many misleading results will be reported and discussed in the future. A general phonetic knowledge is important for the speech researcher with a technical background just as a person working at the cash register is helped by some mathematical skill.

Our own approach at KTH has been to use only broad phonetic labels close to lexical pronunciation in the data bank work. A specific realization is analyzed in acoustical terms rather than in terms of a narrow phonetic transcription. By this approach we have the possibility to study reduction phenomena and context-dependent realizations on a continuous scale. Thus, we will not be dependent on a certain interpretation of a specific linguistic unit.

An important use of speech corpora is the possibility of getting insight into phonemes' variability and context dependence. By search strategies, mentioned in the papers by Boves, Kohler and Carlson and Nord, we can get access to selective samples of realizations. Studies of this type will enrich our understanding of speaker-dependent and speaker-independent habits and our phonetic intuition will be supported by facts. The creation of speech corpora will play an important role in the future phonetic and synthesis research.

Block diagrams and mutual interaction

The task of modelling speech production as part of a text-to-speech system is becoming more and more complex. Traditionally the structure is described by a simple block diagram. Each block is regarded to be relatively independent of the other components. The increased knowledge at each linguistic level demands interaction between different parts of the system and the simple structure has to be replaced be an integrated framework. The source/vocal tract interaction is a typical example of this. In the same way the grapheme-to-phoneme rules have to take into account syntactic information and task-dependent information to a much higher degree than before.

Prosodic models in text-to-speech systems are in the same situation where information from many different sources has to be used. A simple classification of closed and open class words plus simple phrase level rules have taken us a long way, but the current models are more refined. The papers by Fant, Collier, Kohler and Campbell & S.D. Isard discuss these issues in more detail. Information on specific word sequences, semantic load and emphasis has to be included. Methods to predict focal stress as well as rhythmical considerations are needed. With these new models we have or should have tools to take into account extralingistic information as well, supplied by the text source.

Synthesizers

This special issue of the Journal of Phonetics is somewhat limited in the discussion of the sound-generating part of a synthesis system. We find a terminal analog synthesizer in most of the described systems. The fast development of signal processing chips has created a revolution in the possibilities to simulate complex systems. The step back that happened when the first digital systems were designed has now slowly been recovered and the richness in control possibilities is in some cases overwhelming. The need to structure these parameters and a method to do so is discussed in the contribution by Stevens and Bickley. The new generation of terminal analog systems has created new possibilities to simulate alternative voices and to model complex cues. Voice characteristics as discussed in a recent ESCA workshop in Edinburgh, 1990, is an indication of this new trend.

Ultimately an articulatory model will be the most interesting solution for the sound-generating part of text-to-speech systems, when the total flexibility of such a system is appreciated. The development is also going fast in this area, as pointed out in a paper by Fant, but the lack of reliable articulatory data and appropriate control strategies are still some of the bottlenecks. One solution that has attracted interest is the possibility to automatically train neural networks to control such a synthesizer. The paper by Bailly, Laboissière and Schwartz explores such methods that are influenced by control theories.

The most radical solution to the synthesizer problem is to avoid it. Considerable success has been achieved by systems that base the sound generation on concatenation of natural speech units. Sophisticated techniques, like the PSOLA methods, have been developed to manipulate these units, especially with respect to duration and fundamental frequency. Thus, the most important aspects of prosody can be imposed on synthetic speech without considerable loss of quality. However, from the phonetician's point of view this excludes most segmental experimentation, but has been a useful tool for studies of prosodic models as in the paper by Collier. The promise of parametric synthesis as a tool to model and explore all segmental and prosodic aspects of speech and their interactions will drive us to further develop it as a

method in phonetic research. Our ambition is to model natural speech on a more global level, allowing changes of speaker characteristics and speaking style. The use of phonetic knowledge in speech synthesis systems of different kinds is discussed in the paper by Pols and van Bezooijen.

Rules and notations

Development tools for text-to-speech systems have attracted considerable efforts since the computer was introduced into the field of synthesis. It is important to mention the efforts by Holmes, Mattingly and Shearme and the filtered square wave approach by Liljencrants in the development of special-purpose software. The publication in 1968 of The Sound Pattern of English by Chomsky and Halle started a new kind of synthesis system based on rewrite rules. These ideas inspired us to create a special rule compiler for the KTH text-to-speech project in the early seventies. New software is still being developed according to this basic principle, but the implementations vary depending on the developer's taste. It is important to note that crucial decisions often are hidden in the system. The rules might operate rule-by-rule or segment-by-segment. How is the backtrack organized? Can non-linear phonology be used, as in the systems described in this issue by Hertz and Boves? Are the default values in the phoneme library primarily referred to by labels or by features? These questions might seem trivial, but we see many examples of how the design of a system penetrate into the models or even into the thinking of the researcher himself.

Multilingual synthesis

Speech synthesis research of today seems to be in a state of revival. This trend is much due to the increased interest from the speech technology side. This is especially true in Europe where every language deserves a speech synthesis project. Purely industrial reasons would in principle encourage a joint effort to produce truly multilingual solutions. In this way the market can be expanded outside the sometimes minute national markets. The scientific reasons are perhaps even more interesting -- to make cross-language comparisons of languages described in the same framework. These were two important reasons for the development of the KTH multilingual text-to-speech system. Several other projects represented in this issue, from IPO and CNET, for example, have this multilingual flavour. The Boves paper illustrates how a new effort is pursued under the ESPRIT umbrella.

One intriguing project, that is inherently multilingual, is the translating telephony carried out at ATR in Japan. The goal is that speech in a foreign language should be synthesized, ideally reproducing the paralinguistic as well as linguistic content of the original speech.

Skills

We have experienced interesting behaviour in speech researchers. Why are some phoneticians able to develop good rule components for speech synthesis and why are others not able to do this? The answer to this question is still unknown to us. It might be so that the synthesis systems still are so far from good phonetic modeling that special skills are needed to bridge this gap. Another answer might be that the task is so complex and the tools so primitive that the needed simplifications become too restrictive to accept. This special issue will hopefully reject both these explanations, but in the future we need to understand this question in order to better educate speech researchers.

Final remarks

In the past we have seen speech recognition and speech synthesis as rather separate disciplines. At the same time it is obvious that speech production and speech understanding must be studied in the common speech communication perspective. We now perceive a trend towards more commonality in research topics and methods. There are attempts on the higher levels to make components operate in two directions, e.g the two-level morphology that could be used both for production and analysis. On the acoustic level there is an interest in adapting to different speaker and speaking styles in both synthesis and recognition. Adaptive production(synthesis) systems have started to be explored in speech recognition projects. Methods from both synthesis and recognition are integrated in the development of the large speech databases that will be necessary for the advancement of both fields.

A fundamental difference between various speech synthesis researchers is their attitude toward particular methods. We have frequently seen that automatic, self-trained or optimized solutions outperform explicit, knowledge-based approaches. If the objective is increased understanding, the scientist's preference is still the latter, but we must find methods of combining the two traditions. The use of speech databases in combination with rule-based systems is one attempt that we now see emerging.

In this issue we have focussed on speech synthesis in relation to phonetic research. The surge of interest in the commercial application of speech technology can sometimes counteract the free flow of information in research. Undoubtedly the last decades have produced computer-based tools that are widely superior to former laborious experimental facilities. Rule-based, interactive text-to-speech systems connected to flexible synthesizers have made it possible to experiment with global aspects of speech, such as speaker characteristics, speaking styles and dialogue situations. These possibilities have moved the interest away from detailed questions of allophonic realization and coarticulation strategies. The paper by Carlson and Nord, and the paper by Sorin are examples of getting back to these old hunting grounds, using the new tools. If there is a single conclusion to draw from this issue it must be that we still are very far from a complete understanding of most aspects of the speech production process but that we are better equipped than ever to solve the speech code.