An Analysis Method for High Quality Formant Synthesis

P. Badin and G. Murillo
Grenoble, France

1. Introduction

Presently there are in France several laboratories cooperating in a Speech Communication Research Group (GRECO) that is supported by the French National Scientific Research Center (CNRS) and that is working on the constitution of a data base of French sounds. Within this framework, our aim is to carry out an analysis on a corpus of French sounds, in order to generate high quality synthetic speech.

Nowadays, a relatively large number of commercially available systems produce very intelligible speech. However, even if this speech is quite easy to understand, it is not very difficult to detect its synthetic nature. Moreover, it becomes more and more clear that the quality attained is not much improving anymore.

The great progress realized up to now in synthetic speech quality has been essentially based upon research using perception criteria: the acoustic cues contributing to sound perception have been determined exclusively from test verifications of a priori hypotheses. The famous 'locus' theory developed at Haskins Laboratories (Delattre, 1958), for example, was elaborated from perception tests of synthetic speech rather than from objective measures on the loci's values. Even if the synthetic sounds generated by the Pattern Play-Back were mediocre, this approach permitted considerable progress in phonetic knowledge. In fact, this work had been principally focused to determine the cues which allow us to perceive one sound as phonetically different from another one, regardless of its relation to production phenomena.

Since the elements contributing to speech quality are highly complex, and because the equipment employed was rather rudimentary, we consider that this approach is not adequate for research on speech quality and naturalness. Besides, the properties of sound perception are not yet mastered: because of this, perceptual compensation or masking effects caused that some acoustic cues are not important for intelligibility - but that might be important for quality - could be disregarded. Moreover, in order to easily manipulate the synthetic sounds, it is very important to possess reliable references: a production model and the results of an objective analysis.

We believe that, in order to generate very high quality synthetic speech, we

must propose a more fundamental approach which takes into account the speech 'production' aspect, since this is the only way to point out all the acoustic details that might be important from a perceptual point of view. We define a production model and we go back to the classical Analysis-by-Synthesis method proposed for the first time by Bell et al. (1961), but we use more elaborate tools for the analysis and for the comparison as well as for the synthesizer.

2. Method

1. Analysis-Synthesis method principle

The method's goal is to carry out an accurate analysis of natural speech, using a production model. We shall distinguish two levels: the structure of the procedure (i.e. the algorithm surveying the set of operations) and the strategy for its use.

a. Analysis-Synthesis structure

We use the classical Analysis-by-Synthesis scheme: we aim to determine the evolution of the production model's control parameters that will permit us to obtain synthetic speech as close as possible to the original.

The algorithm is divided into two steps:

1. The first one consists of an automatic analysis of the original signal. Poles, bandwidths, F_0 and signal energy are computed. Besides that, sonagrams and DFT of the signal are drawn; finally a graphic display of the speech waveform is also made available;
2. The second step is a feedback procedure: starting from the data acquired in step 1, the evolution of the synthesizer control parameters is determined; the synthetic waveform is computed and compared to the original in order to edit again the control parameters. This procedure is repeated until a correct result is obtained. The comparison is threefold: time, spectrum and perception-wise. Parameter acquisition and editing are done by means of an interactive graphic program.

b. Strategy

It is clear that parameter acquisition and correction cannot be done for all parameters in one single step because of their large number. Once we have acquired the basic parameters (energy, F_0, formants and bandwidths), a first synthetic waveform is computed. This signal may be redrawn on outline, based on data in the literature, in order to get a first approximate result. Immediately after, the stationary zones are refined and verified using mostly the perceptual method (see below). After this, transitions are refined using all helpful analysis data, and proceeding by linear interpolation between the values of the zones surrounding the transition region, every time that analysis results are blurry. For each of these operations the basic parameters

(energy, F_0, formant frequencies) are adjusted first. The rest of the control parameters will serve to refine the results. At this moment, the whole utterance is checked by ear for verification: if the result is not satisfactory, the faulty segments are searched by the perceptual comparison method and readjusted until the whole utterance is considered correct. A long experience in applying this Analysis-Synthesis method shall permit us to increase the performance of the methodology we have described.

2. Tools employed

The production model we have chosen is a parallel type formant synthesizer with a mixed source (periodic signal and noise source) where a 19 parameter-updating is done every 5 ms (this structure is derived from the synthesizer by Klatt in 1980). We have opted for the formant configuration versus LPC synthesis because the former provides a direct acoustic interpretation of the control parameters. This technique, together with DFT spectrum analysis and sonagrams, makes is easier to edit the synthesis parameters by hand.

For the first evaluation of formants we decided to carry out an LPC analysis by the autocorrelation method (Markel and Gray, 1976). Pole and bandwidth values are obtained from the predictor coefficients of the analysis model. Even if this method is not highly accurate, particularly concerning bandwidths, it has the advantage of being fast and completely automatic. The quality of the results is good enough to provide the raw data for control parameter determination.

In an Analysis-by-Synthesis-like method, a most important point is the one dealing with the original versus synthetic comparison: we use a threefold criterion for this comparison. The first one is spectral matching. The second one is a comparison in the time domain : the waveform is displayed on a graphics screen and thus the transition zone boundaries of certain parameters such as noise source energy or voiced source spectrum are determined. The third criterion is the most important one: it takes care of the perceptual comparison ; as proposed by Holmes (1979), it consists of a 'repeated listening to natural and synthetic speech in immediate succession'. The synthetic sound may be composed of a complete synthetic utterance or of a 'synthetically patched natural utterance' (i.e. the natural utterance in which a certain section is replaced by the homologous synthetic one). In this way, it can be determined if there is an unsatisfactory section. This procedure enables one also to locate the different defects that might appear when an utterance is listened to as a whole.

3. First results - Discussion

We have begun to apply this method to the synthesis of a certain number of CVCVC sounds containing French voiced fricatives and stops. About twenty persons listened to a binary comparison-based preference test in which

natural, formant synthesis and LPC synthesis homologous utterances are presented for comparison. The results confirm the high formant synthesis quality and the relevance and efficacy of the method.

The method's exploitation remains at present laborious for the operator: work is in progress in order to make a more interactive system and - more than anything else - to free the operator from jobs not requiring decision taking. The problem is to find a compromise between the operator's decision freedom and the system exploitation heaviness.

This method will further permit a build-up of a dictionary of sounds of a language, and will provide an efficient tool for the determination of acoustic cues of speech.

References

Bell, C.G., Fujisaki, H., Heinz, J.M., Stevens, K.N., and House, A.S. (1961). Reduction of Speech Spectra by Analysis-by-Synthesis Techniques. *J. Acoust. Soc. Am.*, 33, 1725-1736.

Delattre, P.C. (1958). Les indices acoustiques de la parole: premier rapport. (Acoustic Cues of Speech: First Report.) *Phonetica* 2, 108-118 and 226-251.

Holmes, J.N. (1979). Synthesis of Natural Sounding Speech using a Formant Synthesizer. In: *Frontiers of Speech Communication Research*, 275-285. Ed. by Lindblom, B. and Ohman, S. Academic Press, London.

Klatt, D.H. (1980). Software for a Cascade/Parallel Formant Synthesizer. *J. Acoust. Soc. Am.*, 67, 971-996.

Markel, J.D. and Gray, A.H. (1976). *Linear Prediction of Speech*. Springer Verlag, Berlin.