COMPUTER RESYNTHESIS OF SPEECH ON PHONETIC PRINCIPLES

L. Nordstrand and S.E.G. Öhman
Dept of Linguistics, Uppsala

Introduction

In Zetterlund, Nordstrand and Engstrand (1978) and Engstrand, Nordstrand and Zetterlund (1979) it was demonstrated how the prosody (fundamental frequency contour and physical segment durations) of a tape recording of spoken utterances can be manipulated using LPC methods without much affecting the voice characteristics of the speaker or the phonetic identity of the so-called 'segmental' features of the utterance[1]. Using this method it has been possible to perform certain phonetic experiments that lend support to the theory that intonation in Swedish (and probably in many other languages also) is determined by certain (relative) pitch levels that must be reached simultaneously with certain definite nonprosodic sound effects (primarily the vowels) produced during the course of the speech act. On this assumption it appears to be possible to predict the durations of the physical segments from a knowledge of 1) the sequence of phonemes[2] and 2) the relative pitch levels of the phonemes that carry such levels. This should be a basic principle - in so far it holds - for any synthesis-by-rule system.

To carry the phonetic experiments further one should now like to be able to manipulate the parameters of the 'segmental' sound effects, also. I.e. , one should like, in a given recording of a certain utterance, to be able to change the sense-discriminating features[3] of certain occurrences, in that utterance, of various segmental phonemes. It would of course also be desirable to be able to manipulate other kinds of feature, e.g. expressive features, and, in fact, any kind of phonetically relevant feature.

It is not at all obvious, however, how segmental parameters can be controlled in an LPC analysis/synthesis system. The problem can be solved on certain conditions, however. *First*, one has to find a one-to-one transformation from the filter coefficients to a set of parameters whose relation to the synthesized acoustic phonetic effect is explicitly known. Second, in order to make the method easy to use, there should be included a process of interpolation along the time axis between given phonetic sound effects in a phoneme string. Third, to insure that the intermediate sound effects generated by this interpolation procedure do not strike the ear as independent 'intruding phonemes', yet another one-to-one transformation has to be found, one that maps the filter coefficients into a further parameter set.

What parameters to control

Theory as well as experience with formant synthesis motivate us to use spectral parameters for control and manipulation purposes. The transformation that maps the LPC coefficients on the spectral domain is defined by the fundamental theorem of algebra, relating coefficients and roots of a polynomial equation. LPC formants and bandwidths are defined through the roots of the complex polynomial whose coefficients are equal to the LPC coefficients. Efficient algorithms to solve for the roots of the polynomials are available in standard computer systems, and the converse computation of coefficients from roots can be performed by means of a simple well-known formula.

How to interpolate

Any interpolation method to be used must meet two requirements:

(1) it must guarantee that the synthesis filters are stable,

(2) it must not introduce 'intruding phonemes'.

The parameter set that first comes to mind as a possible candidate for these purposes is the set of filter coefficients themselves, i.e. , the LPC coefficients, as they are also called. Linear interpolation of these parameters does not guarantee stability, however [Markel & Gray (1976)] . On the other hand, linear interpolation of the roots (i.e. formant frequencies and bandwidths) of the LPC polynomial always guarantees stability, but this method entails a sorting problem (the so-called formant tracking problem) since the roots in question do not possess any obvious partial ordering. A number of other methods have recently been worked out (see e.g. Markel

& Gray -76), and the parameter sets defined by these transformations are tabulated in Table 1 where data relating to stability are also given. Fig 1 shows how the poles move in the z-plane under linear interpolation of the various parameter sets (in each case using six filter coefficients).

On the basis of the above discussion and informal experiments linear interpolation has been tried using area coefficients, log area coefficients and arc sine of the reflexion coefficients. It was found that synthesis based on interpolation of the arc sine of the reflexion coefficients gave a quality superior to the other methods. Probably this is due to the fact that the arc sine transformation produces a more even distribution of the reflexion coefficients near their peak values, i.e. , when these values are close to 1.

Using this method the following phonetic experiment was performed. The phonetically crucial instants of time in a given utterance was marked with help of a display of the pressure/time wave form. These moments were picked in accordance with phonetic principles sketched in Öhman et al. (1979). LPC coefficients were computed at the crucial instants and stored along with the relevant time coordinate. With this as input data the gaps between the crucial instants were filled with interpolated LPC coefficient sets, so that a new speech wave could be resynthesized.

The signal thus obtained sounded suprisingly similar to the original. One will recognize the voice of the speaker, his dialect, and, of course, what he says without much more distortion than a simple LPC analysis-and-resynthesis (without intervening parameter manipulations) will cause.

Recordings of these tests were played at the conference (see Fig 2).

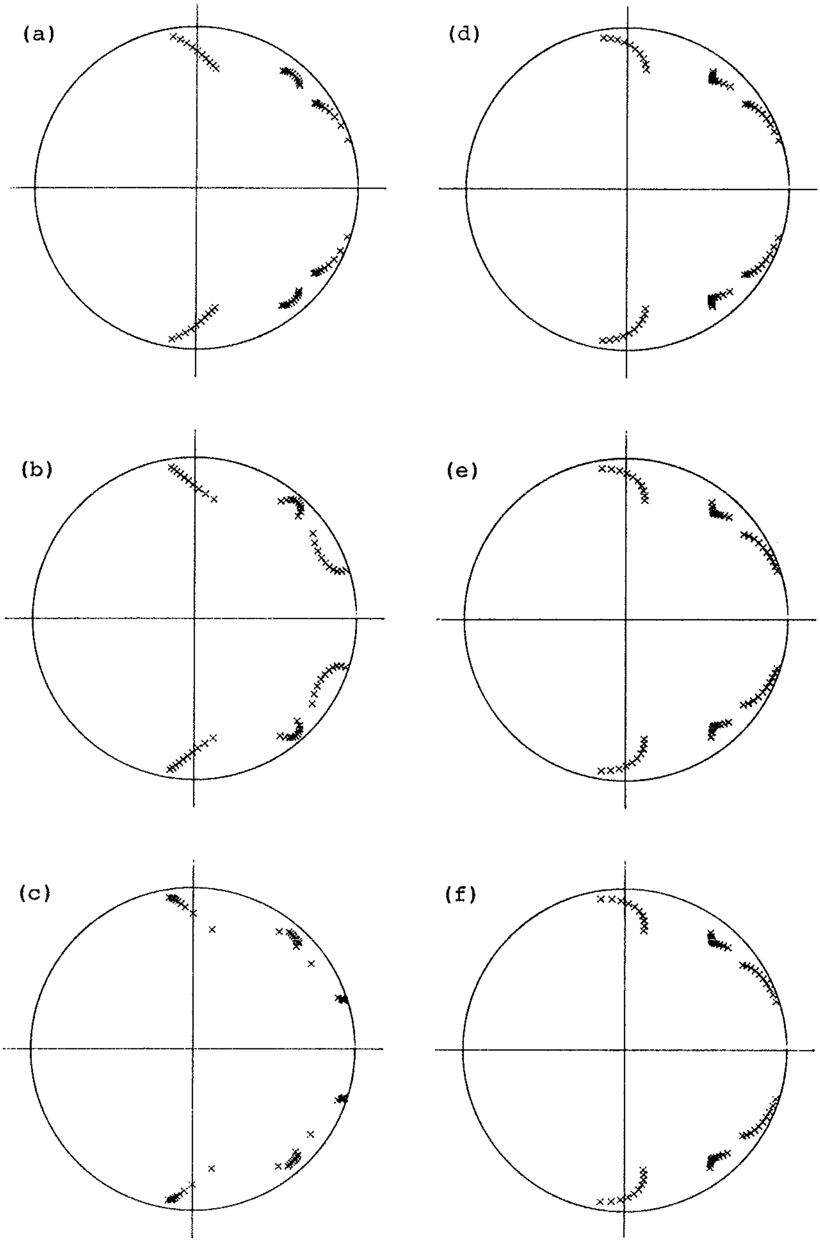

Fig 1. Variations of the root locations (the poles) of an LPC polynomial as a function of linear interpolation of a) filtercoefficients, b) cepstral coefficients, c) autocorrelation, d) reflection coefficients, e) area functions and f) arcsin of reflection coefficients.

Table 1. The various parameter sets based upon the LPC formulation.

Name	Stability
filter coefficients	No
cepstral coefficients	Yes
autocorrelation	Yes
reflection coefficients	Yes
area functions	Yes
arcsin of reflection coeficients	Yes

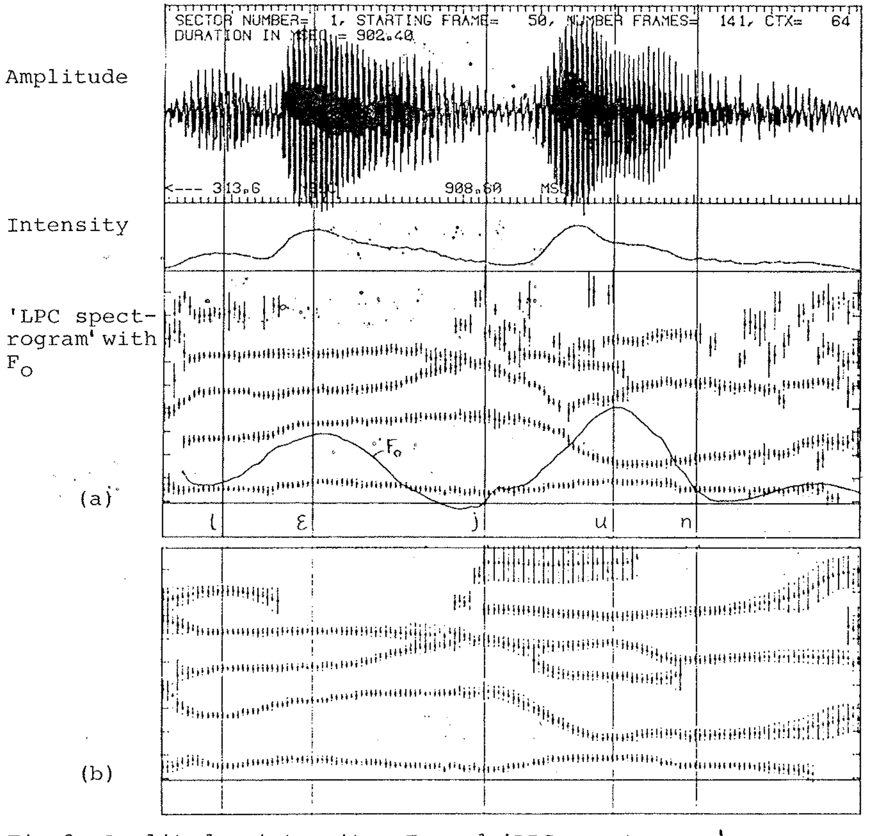

Fig 2. Amplitude, intensity, F_o and 'LPC spectrogram' as a function of time, of the word "lejon" (lion) in the test utterance. a) analysed 'LPC spectrogram', b) interpolated 'LPC spectrogram'. The 'crucial instants of time' are marked with vertical lines.

Notes

1. It is this process of analysis followed by manipulation of analysis parameters followed by synthesis from the parapeters thus obtained that is referred to by the word Resynthesis of the title of this paper.
2. The concept of phoneme intended here is sketched briefly in Öhman, Zetterlund, Nordstrand and Engstrand (1979).
3. Cf. Jakobson & Waugh (1979) for this term.

References

Engstrand, O., L. Nordstrand and S. Zetterlund.1979. Some observations on the role of prosodic parameters in the perception of phrase structure in Swedish. Paper given at The Swedish-Danish Symposium on Phonetics, Speech Transmission Laboratory, Dept of Speech Comunication, KTH, Stockholm Sep. 1979

Jakobson, R. and Linda R. Waugh.1979. The Sound Shape of Language. Harvester Press,Sussex

Markel, J.D. and A.H. Gray Jr.1976. Linear Prediction of Speech. Springer-Verlag, New York

Zetterlund, S., L. Nordstrand and O. Engstrand.1978. An Experiment on the Perceptual Evalution of Prosodic Parameters for Phrase Structure Decision in Swedish. Paper given at The Symposium on the Prosody of the Nordic Languages, Phonetic Laboratory, Dept of general Linguistics, Lund University, June 14-16,1978

Öhman, S.E.G., S. Zetterlund, L. Nordstrand and O. Engstrand. 1979. Predicting segment durations in terms of a gesture theory of speech production. In Proceedings of the 9th International Congress of Phonetic Sciences held in Copenhagen Aug. 6-11,1979. Vol II