A TOOL FOR THE COMPLETE PRODUCTION OF COPY SYNTHESES FROM NATURAL TOKENS

A.M.Simpson
Department of Phonetics and Linguistics, University College London

ABSTRACT

An X Windows graphical user interface for the Klatt Cascade-Parallel Formant Synthesiser [1] is described. It includes facilities for initial synthesiser parameter estimation, for editing time-varying parameter values, and for aural, spectral, and spectrographic comparison of target and copy-synthesised stimuli.

COPY SYNTHESIS

High-quality copy synthesis allows the creation of speech-like stimuli which are sufficiently natural to ensure listeners are listening in the speech mode. The stimuli can be altered to investigate the relative contribution of different acoustic cues in encoding phonetic contrasts.

Formant synthesis produces speech-like stimuli by passing a source waveform through a complex filter whose resonances model those of the vocal tract. The precise nature of the source can be varied (e.g. periodicity, voice fundamental frequency, open quotient, jitter) as can the characteristics of the filter (e.g. formant frequencies, bandwidths, and amplitudes, and the presence of anti-resonances). This flexibility allows the complex spectro-temporal variation that occurs in natural speech to be closely modelled.

However, the specification of formant synthesiser parameter values is a complex and laborious process. Even with good initial estimates of the voice fundamental frequency contour and the formant trajectories, much work still has to be undertaken to model the variation in formant amplitudes and bandwidths before a synthetic copy will sound natural.

The task of refining synthesiser parameter values is made difficult by the inability to visualise how synthesiser parameters co-vary, to edit them easily, and to assess easily the effects of any manipulations both aurally and by using more objective analysis methods.

This tool addresses such difficulties by providing all such facilities within an integrated package which includes a version of the Klatt Cascade/Parallel Formant Synthesiser. It enables users to calculate initial parameter values from the target stimulus, to edit such values easily, to see and hear both natural target and its synthetic copy, and to perform spectral and spectrographic analysis of both to ensure closeness of match.

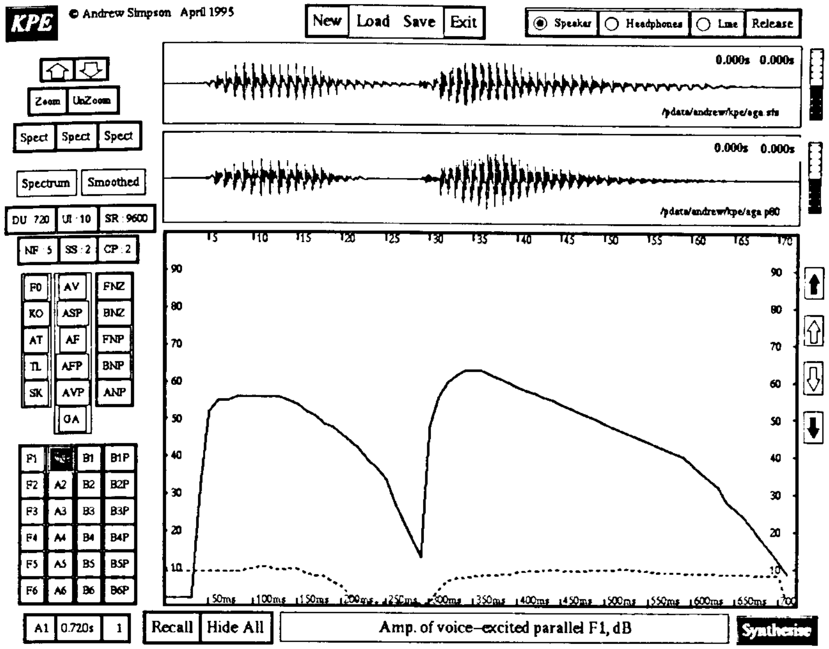
INITIAL PARAMETER ESTIMATION

The tool includes the facility to calculate the voice fundamental contour from the target natural token. In addition, it is possible to specify formant frequency trajectories by tracing each formant's path onto a spectrogram

PARAMETER EDITING

Time-varying parameters are edited using a 'canvas' onto which each parameter's trajectory can be drawn using a mouse-controlled cursor. Any number of parameters can be simultaneously displayed to allow users to co-ordinate the value of parameters which vary time-synchronously, for example, the formant amplitudes at the onset of voicing after plosive release, or at a plosive's release burst. Each parameter is displayed using a different colour to ensure it is distinguishable from others; parameters can be cycled through three states: being edited, displayed only, and not displayed, allowing arbitrary groups of parameters to be simultaneously displayed and edited. Figure 1 shows the amplitude of the first formant being edited, whilst its frequency trajectory is also displayed.

It is possible to specify parameter values with great accuracy as the precise parameter value at the cursor is displayed. The parameter canvas is time-aligned with both target and copy-synthesised waveforms and vertical cursors indicating the point in time being considered are displayed in all windows, allowing alignment between the two waveforms, and allowing the monitoring of the effect parameter changes have on the synthetic waveform. A parameter's trajectory can either be specified by drawing a line between start and end points, or can be drawn free-hand. To facilitate inspection and editing of parameter values over very brief regions of the stimuli a zoom facility allows users to display a region in greater detail, allowing very brief acoustic cues such as release bursts to be inspected and specified with great accuracy. Figure 2 shows how the zoom facility has been used to display the burst and first few cycles after release of an intervocalic voice plosive.



DU 720 UI 10 SR 9600
NF 5 SS 2 CP 2
A1 0.720s 1
Amp. of voice–excited parallel F1, dB

Figure 1. Natural (upper) and synthetic waveforms of /aga/ together with the values of A1 (amplitude of voice-excited parallel first formant, solid line) and F1 (first formant frequency, dotted line). Any number of the time-varying parameters can be displayed.

COMPARISONS

Aural comparison of complete or partial target and copy-synthesised waveforms is possible simply by marking the extent of the desired region using mouse-controlled cursors and then clicking on the appropriate waveform. It is also possible to calculate the amplitude spectrum of corresponding regions of both waveforms thus providing more detailed information about their degree of similarity. This is illustrated in Figure 2 where the release burst spectra of a natural and synthetic intervocalic voiced velar plosive are compared. Although the strong burst at around 2 KHz has been modelled well, there are discrepancies around the secondary peak at 4 KHz and in the region below about 200 Hz. Such comparisons are useful for comparing short or relatively unchanging regions of the signal, or for making gross spectral comparisons

To assess how closely the complex spectro-temporal variation seen in speech has been modelled a spectrogram calculation facility is provided which can

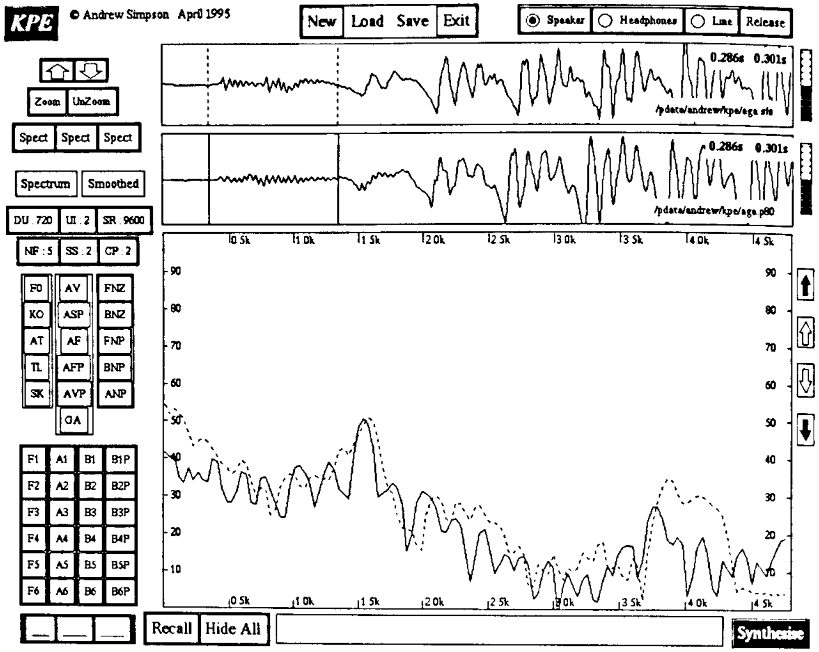


Figure 2. Detail of the burst and few cycles after release for natural (upper) and synthetic /aga/ stimuli. The amplitude spectra for the marked burst regions are compared below.

display either natural or synthetic waveforms, or both simultaneously, aligned in time to facilitate comparison.

FILE FORMATS

The tool supports a range of audio data file formats including Sun Audio format (.au), Microsoft audio format (.wav), and the SFS [3] format (.sfs). The synthesiser parameters are stored in the form of an ASCII text file with each parameter's value for each successive frame of the stimulus stored as a number.

PORTABILITY

The tool has been written in C using SUIT [2] and requires a UNIX platform with X Windows. Versions currently exist for Sun Sparc and Linux/XFree86 architectures. Plans exist to port the tool to other platforms. Contact the author (*andrew@phon.ucl.ac.uk*) for more details and for information about how to obtain the tool.

ACKNOWLEDGEMENTS

The formant synthesiser is an implementation of the Klatt Cascade-Parallel Formant Speech Synthesiser by Jon Iles (*j.p.iles@cs.bham.ac.uk*) and Nick Ing-Simmons (*nicki@lobby.ti.com*) The spectral, spectrographic, and pitch-extraction are from SFS [3].

REFERENCES

[1] Klatt, D.H. (1980), "Software for a cascade/parallel formant synthesiser", *Journal of the Acoustical Society of America*, vol. 67(3), pp. 971-995.
[2] SUIT, The Simple User Interface Toolkit, University of Virginia, (*suit@uvacs.cs.Virginia.EDU*).
[3] The Speech Filing System, Dept of Phonetics and Linguistics, University College London. (*sfs@phon.ucl.ac.uk*)