LACS: LABEL ASSISTED COPY SYNTHESIS

M. Scheffers and A. Simpson
IPDS, Kiel, Germany

ABSTRACT

We describe a knowledge-based method of deriving the control signals for the Klatt synthesizer. This method uses a combination of a rich acoustic analysis and an intelligent post-processing system which employs labels from the segmentation of the original signals to modify and augment the analysis data.

INTRODUCTION

The automatic generation of control signals to a drive a formant synthesizer offers an excellent method of validating phonological models by observing their phonetic output. This is made all the more challenging by the high quality of the speech which formant synthesis can produce when provided with appropriate control signals.

A synthesizer such as the early Klatt model [1] offers a large number of control parameters allowing adequate modelling of the acoustic products of the vocal tract.

However, obtaining parametric values which are to serve as the phonetic correlates of the phonological systems and structures of a language is a laborious task. One of the most interesting and enlightening methods of arriving at these numbers is undoubtedly copy synthesis [2], i.e. driving a synthesizer with the results of the analysis of a natural utterance.

There are, however, two serious problems involved in mapping the results of an acoustic analysis onto the control parameters of the Klatt formant synthesizer.

First, there is a discrepancy between the information delivered by the acoustic analysis of an utterance and the rich variety of synthesizer parameters which can be used to model the acoustic signal. Most acoustic analyses, for example, only allow a decision to be made as to whether the resonators should be excited with a periodic (Fo found) or an aperiodic source (no Fo found). The Klatt synthesizer, on the other hand, offers four dynamic parameters which can be used to model glottal and supraglottal sources, two parameters to model the aperiodic (glottal and supraglottal friction) and two for the periodic source (voicing and a low-pass filter to model the voicing in voiced stops and fricatives).

Second, parametric information about more complex products of the vocal tract is usually not available in the analysis. Voiced fricatives are an example of this. A voiced fricative such as [z] leaves an analysis either as a voiceless fricative (no Fo found) or as a frictionless approximant (Fo found). Although the former may be the most appropriate analytical outcome for a synthetic utterance, neither allows the original fricative to be modelled. Breathy voice presents a similar problem.

In this paper we would like to describe a method which attempts to overcome these problems by subjecting analysis data to an intelligent post-processing based on the manual segmentation and labelling of the original signals.

An acoustic data base of read speech such as that constructed at the IPDS Kiel [3] provides an excellent source of segmented and labelled signals. This data base contains 31374 segmented and labelled words of German spoken prose. Labels are primarily phonological in nature. Each phonological label is time-aligned with the start of a signal portion representing the chief phonetic correlates of the phonological item in question. Phonological labels are supplemented by quasi-phonetic labels to indicate aspects such as creaky voice, plosive release phases and vowel nasalization when other correlates of a nasal are absent.

ANALYSIS

We begin by describing the analysis method.

The short-term energy (RMS), Fo and formant analysis facilities of the ASSP signal processing package developed at Kiel [4] are used to obtain initial estimates of the parameters for the Klatt synthesizer. Fo values are analysed using an extremely fast and highly accurate periodicity detector [5,6]. Formant frequencies and bandwidths are determined by root-solving of the LPC polynomial [7,8]. Subsequently, formant amplitudes are estimated from the LPC spectrum. The speech signals being sampled at 16 kHz, 8 formants are analysed for male voices.

Conditioning and Sorting

Because the formant analysis always provides the number of formants specified, it must introduce pseudo-formants when there are fewer resonances. One class of pseudo-formants results from real roots. These receive a fixed, very high or low frequency and/or an extremely large bandwidth. The other pseudo-formants are also characterized by a very large bandwidth and occur either about midway between two "true" formants or very near to one.

Whereas these pseudo-formants are accommodated for in the synthesis model applied in ASSP, the discontinuities they cause in the formant tracks are disruptive in the Klatt model and hamper (semi-)automatic processing and interpretation of the data. The raw data are therefore sent through a conditioning stage: *ksort*.

First, the pseudo-formants resulting from real roots are removed. The bandwidth of the other formants is checked against a threshold, currently set at 1000 Hz. If it is above threshold, a set of heuristic rules is invoked to decide whether the formant is to be deleted or to be merged with a nearby formant (weighted mean).

Next, each formant is assigned a best fitting formant number by comparing its frequency with a list of average formant frequency values. When two formants receive the same number a more global best match is searched in which as few formants as possible obtain numbers different from the ones originally assigned to them.

Finally, gaps in the resulting formant tracks are filled with dummy values which can easily be identified by the next processing stage.

Analysis Results

The Fo analysis lives up to its reputation: in the 100 sentences currently under study, no gross errors were found. The few errors made mainly consist of:

- delayed voicing detection due to irregularities in the initial glottal pulses,
- failure to detect creaky voice,
- failure to detect stretches of weak and noisy voicing often found in utterance final syllables.

If these Fo errors are found to detract seriously from the quality of the synthetic utterance, as can happen at voiced-voiceless boundaries, they can easily be manually corrected. Failure to detect creaky voice is an exception to this and is one of the areas where label information can be successfully used for automatic correction (see below).

For voiced sounds, the lower formants are generally consistently found and numbered correctly. Keeping in mind that we only need the lower four formants for these sounds, these data can directly be used in the synthesis. Exceptions are typically nasals and nasalized vowels, where an additional nasal formant at about 2.5 kHz is detected. For nasals, this presents few problems since the discontinuities are aligned with the nasal closure and release. In fact, nasals come out quite nicely in the synthesis. Nasalized vowels pose a bigger problem because the formant sorting goes awry.

For unvoiced sounds, there are more diverse problems. First of all F1 data are rarely found and in many cases F2 data are also absent. Second, the scatter often found in the formant data that are present makes it difficult to properly number the formants. Although the absence of lower

formants may seem to pose no problems because the corresponding resonators are not excited in the synthesis, the Klatt model does use their frequency values to adjust the amplitudes for the higher ones.

We have recently started experimenting with a different kind of analysis, the so-called 'Robust Formant Analysis' [9]. As with root-solving, it delivers the number of formants specified, but the formant tracks are virtually continuous. Some other properties of the data obtained by this analysis are:

- F1 is continuously present and has a reasonable course.
- F2 corresponds quite closely to the values found by root-solving or peak-picking.
- In closures the data tend towards those of an open tube rather than scatter as in the other analyses.

However, since formants are defined purely operationally in this analysis and need not correspond to resonances in the spectrum, we observed that especially in the mid-frequency region (roughly 2 to 4 kHz) resonances are often represented by two "formants". Since their frequencies are rarely close, it is nearly impossible to detect this and merge the data. Presently, we are looking for ways to combine the results of the two analyses using the strength of each to compensate for their respective weaknesses.

POST-PROCESSING

The first pass through the data delivered by *ksort* ensures that any gaps in the formant tracks are filled in as harmless a fashion as possible. In general, this entails nothing more than carrying out a simple interpolation between two formant values. Furthermore, normalization of the RMS values is carried out and formant amplitudes are modified to compensate for the corrections made in the Klatt model.

Next, the analysis data are combined with labels from the manual segmentation such that each analysis frame is associated with one label. In certain cases the labels which are in sequence in the data base are collapsed into one. So, for instance, creaky vowels are represented in the data base as a sequence of two labels, the first of which indicates the presence of creak over the following vowel. So that this information does not get lost the vowel label is suffixed with a creak marker.

The second pass through the data uses the information provided by the labels to map the analysis data onto synthesizer control parameters. Below, we present three examples of the way in which label and analysis information can be successfully combined to exploit to the full the control parameters made available in the synthesis model. All the examples deal with different aspects of mapping analysis data onto control parameters for the source signals.

Creaky Voice

Portions of creaky voice are generally not found in the Fo analysis. If the label information indicates that a vowel is creaked, but frames have been declared unvoiced in the Fo analysis, creak is modelled by inserting random low Fo values from the vowel onset until the first voiced frame is found. Although it is not possible to model many aspects of creaky voice in the Klatt model we are using, such Fo values together with the fluctuating amplitude of voicing derived from the RMS values produce perceptually acceptable creaky voice.

h and its Correlates

Voiceless signal portions annotated with **h** are assumed to represent periods of turbulent airflow originating at the glottis. These are modelled by mapping the RMS value onto the amplitude of aspiration. Frames labelled with **h** and returned as voiced from the Fo analysis are considered to be periods of breathy voice. The RMS value is used to set the amplitude of voicing. Following [1,10], values for the parameters for the amplitude of aspiration and the amplitude of sinusoidal voicing are derived by subtracting 3 dB and 6 dB, respectively, from the voicing value.

Plosive release and aspiration

The label **-h** annotates a signal portion from the plosive release to the end of any aspiration. The place of articulation of the release is derived from the preceding plosive label and the following vowel. The release is modelled by using RMS values to set amplitude values for the supraglottal fricative source. Once the burst and initial release phases have passed, the RMS values are mapped onto the aspiration source. The length of the supraglottal and glottal friction are varied with the place of articulation of the plosive, the local friction being maintained longest for dorsal plosives.

DISCUSSION

The main aim of the copy synthesis method described here is to derive parameters for rule-driven synthesis. Using phonological/phonetic information allows us to be very selective in the way in which we modify the analysis data to arrive at synthesis parameters.

The advantages of this approach are manifold. Analysing, processing and synthesizing a large number of utterances is fast. Auditory inspection quickly identifies naturally sounding stretches of synthetic utterance. These are places where we can assume that the parameter courses can be used to derive the correlates for the rule-driven synthesis.

The modifications we can carry out on the basis of label information are wide-ranging. They can reflect the findings of others, e.g. the RMS mappings in breathy voice which are based on numbers taken directly from the literature. Other modifications can represent a step-by-step idealization of the analysis data. This is especially desirable when working towards easily definable parameter courses in synthesis-by-rule. The process of idealization can be gradual, allowing the consequences of each step on the naturalness and acceptability of the resulting signals to be assessed. The ultimate modification is to completely discard analysis data for difficult portions, such as voiceless fricatives, and insert numbers obtained elsewhere[1].

The next step in our work will be to investigate the advantages of using label information already at the analysis stage.

ACKNOWLEDGEMENT

We would like to thank Lei Willems for providing us with the software of the robust formant analysis.

REFERENCES

[1] Klatt, D.H. (1980), "Software for a cascade/parallel formant synthesizer," *J. Acoust. Soc. Am.*, vol. 67, pp. 971-995.

[2] Holmes, W.J. (1989), "Copy synthesis of female speech using the JSRU parallel formant synthesizer", *Proc. EUROSPEECH*, vol. 2, pp. 513-516.

[3] IPDS (1994), *CD-ROM#1: The Kiel Corpus of Read Speech*, vol. I, Kiel: IPDS.

[4] Scheffers, M., Thon, W. (1991), "Workstation and signal processing software for experimental phonetics", *Proc. XIIth ICPhS*, vol. 2, pp. 486-489.

[5] Schaefer-Vincent, K. (1982), "Significant points: Pitch period detection as a problem of segmentation", *Phonetica*, vol. 39, pp. 241-253.

[6] Schaefer-Vincent, K. (1983), "Pitch period detection and chaining: Method and evaluation", *Phonetica*, vol. 40, pp. 177-202.

[7] Vogten, L.M. (1983), *Analyse, zuinige codering en resynthese van spraakgeluid*, doctoral thesis, Eindhoven University of Technology.

[8] Saito, S., Nakata, K. (1985), *Fundamentals of speech signal processing*, Tokyo/Orlando/London: Academic Press.

[9] Willems, L.F. (1987), "Robust formant analysis for speech synthesis", *Proc. Eur. Conf. Speech Technology*, vol.1, pp. 250-253.

[10] Allen, J., Hunnicutt, M.S., Klatt, D. (1987), *From text to speech: The MITalk system*, Cambridge: CUP.

---

[1] an idea suggested to us by John Local.