FORMANT DIPHONE PARAMETER EXTRACTION UTILISING A LABELLED SINGLE-SPEAKER DATABASE

Robert H. Mannell

Speech, Hearing and Language Research Centre (SHLRC)
Macquarie University, Sydney, Australia

ABSTRACT

This paper examines a method for formant parameter extraction from a labeled single speaker database for use in a formant-parameter diphone-concatenation speech synthesis system. This procedure commences with an initial formant analysis of the labelled database, which is then used to obtain formant (F1-F5) probability spaces for each phoneme. These probability spaces guide a more careful speaker-specific extraction of formant frequencies. An analysis-by-synthesis procedure is then used to provide best-matching formant intensity and bandwidth parameters. The great majority of the parameters so extracted produce speech which is highly intelligible and which has a voice quality close to the original speaker.

1. INTRODUCTION

Speech synthesis by the concatenation of formant parameter diphones is not often attempted for a number of reasons, most relating to the difficulty of formant parameter extraction and to the reliability of the extracted parameters. Further, the formant model is not a good model of certain consonant classes such as the stops and fricatives. Across such consonants, automatically-extracted formant tracks often appear to be almost random. Erroneous formant parameters have a strong tendency to degrade synthetic speech quality "catastrophically" rather than "gracefully".

Synthesis techniques based upon LPC-parameter or waveform concatenation are much less vulnerable to the effects of poorly extracted parameters. The formant model is, however, more straightforwardly related to the source-filter model and thus to speech production. Whilst it is true that overlap-add concatenation of waveform-based diphones can easily model a voice with quite high fidelity, new voices and voice qualities require the recording of new speakers (or the same speaker utilising a different voice quality) and the extraction of a new diphone database. Such systems can be used to examine the effects of intonation and rhythm on voice quality or vocal affect but formant-based systems can much more readily examine the effect of frequency-domain modifications on voice quality. Such modifications might include formant frequency shifting, bandwidth modification, modification of relative formant intensities and spectral slope variation. It is even possible, if the synthesiser design allows it, to experiment with the insertion of additional poles and zeroes into the spectrum such as might occur when modelling the "singer's formant" for certain styles of singing voice. Such research requires a parallel formant synthesiser with a great deal of flexibility of control. Further, and most importantly, it requires a diphone database that is extremely accurate. Formant errors must be minor and few in number and this should be achieved without excessive hand correction. Formant tracks should display, as far as possible, pole continuity across fricatives, stops and affricates. Extracted intensities and bandwidths, upon resynthesis, should result in spectra that are as close as possible to the original natural spectra.

This paper describes an analysis-by-synthesis method that attempts to achieve the above goals.

2. FORMANT FREQUENCY TRACKING

This algorithm does not represent a simple formant analysis system for the following reasons:-

- the algorithm only works on segmented and labelled speech databases in which all phonemes and some sub-segmental features (including monophthong and diphthong targets) have been identified.
- the algorithm accomplishes its task in several passes rather than in one pass.
- a 14 coefficient and a 24 coefficient LPC are used, rather than a single LPC analysis
- all final decisions are constrained by phonetic expectations
- formant bandwidths and intensities will be extracted utilising an analysis-by-synthesis method

The method described requires a fully segmented and labelled speech database (sampled at 10 kHz) for the target speaker. Formant tracking is achieved by a number of analysis passes.

The formant tracker was written especially for this procedure and is a familiar LPC based formant tracker. Two versions are used, a 14 coefficient and a 24 coefficient LPC. The 14 coefficient LPC is used to determine the position of the major spectral peaks, whilst the 24 pole formant tracker is used to determine accurate peak positions, especially for closely spaced formants which may appear fused in the 14 coefficient analysis.

2.1 First Pass

Stage 1. Stage 1 of the first pass examines the speaker's utterances of the neutral vowel /ɜː/ in order to determine an approximate mean value for each of the speaker's formant frequencies (nb. success may be dialect dependent, but for Australian English this vowel is quite central and is not affected by following post-vocalic /r/). Only /ɜː/ vowels that were not adjacent to a nasal consonant were selected. In this stage of the first pass only the vowel target, as defined by the segmenting and labelling process, was used.

Only the 14 coefficient LPC was used for this pass as formants for /ɜː/ are approximately evenly separated and the resulting spectrum should both resolve all of the formants and be free of spurious peaks. Any peaks below 250 Hz were ignored in this analysis. If there are only four evenly separated peaks for a particular speaker and their spacing (averaged for all /ɜː/ vowels for this speaker) would predict a fifth formant above 5kHz then the subject would henceforth be assumed to have a short (relative to an average adult male) vocal tract and only 4 formants would be determined in all subsequent analyses. Children's speech was not analysed so it was assumed that all speakers would have 4 or 5 formants below the Nyquist frequency (5 kHz).

These approximately neutral formant values (referred to below as "N1" to "N5") and the average spacing between the formants for the neutral vowel ("Ns" below) were used to constrain the formant tracker to select reasonable candidate peaks for all of the non-nasalised vowel targets in stage 2 of the first pass.

Stage 2. Utilising the N1 to N5 values and the Ns value calculated in stage 1, above, a probability space for each vowel class by each formant was separately determined. These probability spaces varied for F1 to F3 depending upon the vowel class that the vowel being analysed belonged to. F4 and F5 had a single probability space each for all vowel classes. The probability space determination utilised both these calculated speaker-specific values and expert knowledge of the relative formant space for Australian English vowels.

For example, different F2 probability formulae were derived for front, central and back vowels, whilst different F1 probability formulae were derived for high, mid and low vowels. (eg. for the Australian English central vowels /ʉː, ɐː, ɐ, ɜː, and ə/ (as well as the first target of /ai, au/ and the second target of /iə, uə/) p=1 IF (N2 - Ns/2) <= F2 <= (N2 + Ns/2), p=0 IF F2 >= N3 and p=0 IF F2 <= N1. This defines a central space with a probability of 1, an outside space with a probability of 0 and an intermediate space where probability varies linearly from 0 to 1).

In stage 2 of the first pass the targets of all non-nasalised monophthongal and diphthongal vowel phoneme targets (as defined by the labels) were analysed for F1 to F5 utilising the probability spaces defined in stage 1, above. Vowels in the context of a nasal consonant were ignored for this pass in order to avoid the difficulty of differentiating between F1 and the nasal formant (henceforth Fn). For this stage of the first pass only the 24 coefficient LPC was used to ensure that all relevant peaks were resolved.

During stage 2 of the first pass no formant trajectory constraints were applied to formant frequency selection. Each vowel target analysis frame was treated entirely independently.

A major difficulty for this, and subsequent stages, was the determination of some heuristic that dealt with a situation where two peaks occurred in a non-zero probability frequency range. Selecting the highest probability peak always occurred when the two peaks were of approximately equal intensity (to within 1 dB). When two peaks had equal probability (to within ±0.1) then the more intense peak was selected. When two peaks had both equal intensity and probability then the peak closest to the mean (N1 to N5) value was selected. A problem arose when a lower probability peak was more intense than a higher probability peak. A heuristic was utilised which deducted p=0.1 from the less intense peak for each 1 dB less intense it was relative to the most intense peak in the non-zero probability region, but this heuristic was only applied when this second lower probability but higher intensity peak was not the highest probability peak for an adjacent formant.

Once all 4 or 5 formant frequencies were selected for each analysed frame of each monophthong and diphthong vowel target within the database, the results were displayed graphically for measured values of each vowel phoneme target independently on F1/F2, F2/F3 and F4/F5 planes. A trained phonetician with expert knowledge of Australian English vowel formant values graphically selected an ellipse that encompassed all target values that were determined to be validly analysed (a number of clearly mis-analysed vowel formant values still remained and these were excluded here). The selected ellipses defined the 0.9 to 1.0 probability space for each vowel target in the second pass.

2.2 Second Pass

The probability spaces determined in stage 2 of the first pass are much more constrained than those determined by stage one of the first pass. The new probability spaces are specific to each individual vowel phoneme (rather than generalised across a number of vowels), and closely represent the actual productions of each vowel target by the speaker being analysed. The inner probability ellipsoid for each target that was defined graphically in stage 2, above, defines the p=0.9 to 1.0 space (0.9 on the ellipsoid boundary and 1.0 in the centre). An outer probability ellipsoid defining probabilities 0.0 to 0.9 is also defined with the same centre and rotation as the inner ellipsoid, but with its boundaries exactly twice as far from the ellipsoid centre as for the central ellipsoid.

In the second pass the 24 coefficient LPC analysis is now applied to the entire database.

Stage 1. Stage one of the second pass first examines all of the vowels in the database including their transitions and between-target glides in the diphthongs. Unlike the first pass, this pass applies both the probability calculations and also trajectory constraints to the determination of formant frequencies. First, all vowel targets are recalculated using the new more constrained second pass probability spaces. Whenever a vowel target has two candidate LPC peaks for any formant that are within p=0.5 of each other, both are temporarily stored. Initially for the targets only, the trajectory through the most probable peaks is determined for each formant for each target. Less probable peaks are then examined to determine if a smoother trajectory can be determined through the LPC poles, but with the constraint that the higher probability peaks must still represent >50% of the points in each formant trajectory.

The formants F1-F4 are reliably tracked using this algorithm but F5 and the nasal formant (Fn) are often missing. That is, there are no candidate poles in the LPC analysis. Missing Fn and F5 are simply skipped (by giving them an impossible value of -999) and are later generated by interpolation from detected values. If a whole phoneme has missing Fn or F5 values then default (p=1) values are allocated for the whole phoneme.

The next step is to track the formants between the targets of each diphthong (or between two adjacent vowels). The probability space for such between-target transitions is determined by linear interpolation between the target probability ellipsoids, as shown in figure 1. Whilst it is obvious that such transitions are rarely perfectly linear, the linearly interpolated transitional probability space appears to assist with the tracking of formants through these transitions with very few errors, especially when transition trajectory constrains similar to those applied to the targets are applied to these transitions.

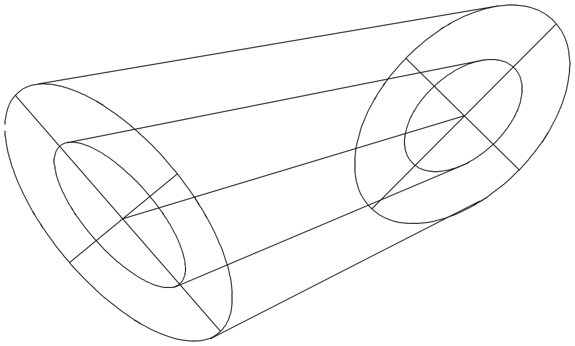


Figure 1: Diphthong 2 target probability ellipses and interpolated transition space (on a two-formant plane).

The next step in stage 1 of the second pass is the tracking of consonant-to-vowel and vowel-to-consonant transitions to complete the formant analysis each entire vowel as defined by manual segmenting and labelling. This stage is one of the most problematic stages as there is no pre-determined offset or onset formant probability space. The onset and offset vowel transitions are tracked outwards from the targets. Transition trajectory constraints and simple formant order constraints (eg. F2 > F1) are the only available formant constraining variables.

Once these rough transitions are calculated the onset and offset values are displayed graphically to a trained phonetician in a manner identical to the procedure utilised in stage 2 of the first pass and a probability ellipsoid is determined graphically in exactly the same way as before to produce an onset/offset probability space for each consonant which will be used in stage 2 of the second pass. (nb. onset and offset values have been conflated into a single probability space).

Stage 2. Utilising the onset/offset consonant probability ellipsoids, as well as transition trajectory constraints, the transitions from the vowel targets to the adjacent consonants are re-calculated. This is followed by the calculation of the consonant target formant frequencies (modelling pole continuity rather than actual front cavity resonance peaks for consonants such as /s/). These calculations utilise the probability spaces for each consonant, as determined for the onsets/offsets, as well as the usual transition trajectory constraints. This procedure proceeds outward from each vowel and analyses progressing from each vowel meet at the centre point between each pair of vowels (or at the utterance start or end points for initial or final phonemes).

At the end of stage 2 the formant frequencies have been completely calculated for all phonemes. All calculated formant tracks are then examined visually and hand corrected when necessary. Such corrections occur most frequently at the point when two analyses proceeding outwards from two vowels meet somewhere within a consonant cluster. Nevertheless, for the single subject whose speech was processed in this manner, less than 5% of tokens required hand correction even for the consonants.

3. FORMANT INTENSITY AND BANDWIDTH CALCULATIONS

This phase of formant parameter analysis is necessary as the parallel formant synthesiser to be utilised for resynthesis of the speech from concatenated formant-parameter diphones requires not only formant frequency values, but also formant gain and bandwidth parameters. (This synthesiser is the SHLRC Parallel Formant Synthesiser, "MU-TALK", which is a software version of the hardware synthesiser described in Clark et al. (1986) and Summerfield and Clark(1986)).

An analysis-by-synthesis methodology is used to determine the intensity and bandwidth parameters which, upon resynthesis, result in the smallest frame-by-frame Euclidean spectral distances between resynthesised and the original speech spectra. This analysis-by-synthesis methodology utilises the same synthesiser that will be used in the target text-to-speech system. This is essential to ensures that any extracted gains and bandwidths are meaningfully related to the synthesis hardware to be used.

Initial values for the formant bandwidths are set at reasonable first estimates of the bandwidth of voiced or voiceless speech sounds for each formant centre frequency.

$$Bx = (80 + 120 \times Fx / 5000) \times W$$

Where W=1 for voiced speech and W=2 for voiceless speech.

Initial intensity values are set to be equal for all formants and equivalent to the RMS average intensity of each speech frame being analysed. Formant intensities are varied in steps of 1 dB, which corresponds approximately with many estimates and measurements (eg. Florentine et al., 1987) of intensity difference limens in the frequency range 200-5000 Hz.

The initial analysis assumption is that the bandwidths are correct. This permits the gains to be varied without the bandwidths independently interacting with them to also vary peak intensity. For the purposes of this algorithm, formants greater than 0.8 x Ns (the average formant separation for /ɜː/, see section 2.1, above) apart are assumed to be independent of each other and their intensities are varied without worrying about the effect of each upon the other's peak intensity. Formants that are less than 0.8 x Ns apart are assumed to affect each other's peak intensity and so their intensities are varied as a pair. The effect of this assumption is that F1 and F2 of the back vowels, F2 and F3 of high front vowels, and Fn and F1 of all vowels, need to be analysed together. For example, a change in the gain of F2 for /i:/ affects the intensity of F3 when /i:/ is resynthesised. The analysis-resynthesis algorithm is applied one formant (or one pair of formants when < 0.8 x Ns apart) at a time and intensities are modified until the smallest Euclidean distance is determined between natural and synthetic spectra around the formant being analysed. The spectral regions being analysed for spectral distance around each formant are the frequency bands centred upon that formant and extending to the mid-point between that formant and the adjacent formant (or to 0 Hz for the lower bounds of Fn, or 5000 Hz for the upper bounds of the highest formant). After each formant intensity modification the frame is resynthesised and tested for spectral distance from the original natural speech frame. Spectral distance measures are carried out on smooth 24 coefficient dB-scaled LPC spectra.

When all formant intensities have been determined, the bandwidths are modified (without further modifications to the formant gains) until the smallest spectral distance is achieved. Bandwidth modification is in steps of 20% of each preceding bandwidth, which Flanagan (1972) suggests is approximately the formant bandwidth difference limens.

Because the spectral distance measures examine absolute intensity differences as well as differences in spectral shape, the original segmental and supra-segmental intensity profiles, as well as the relative formant intensities which define spectrum shape, are all recovered by this analysis method.

4. DISCUSSION

When the procedure described above is applied to a corpus of single-speaker speech data, there are less than 5% formant tracking errors and very few obvious intensity and bandwidth analysis errors. When source and F0 information for a target sentence are also carefully analysed and this information is combined with analysed formant frequency, intensity and bandwidth data, and the speech is resynthesised, the quality of the original and re-synthesised sentences are quite close. A sample natural sentence [SOUND 0627_01.WAV] and the resynthesised sentence [SOUND 0627_02.WAV] are provided on the cd-rom version of this paper for comparison.

The procedure outlined above is a very computationally intensive procedure that also requires some operator intervention, most importantly during the phases where outliers are excluded during the determination of formant frequency probability spaces. Another disadvantage of this algorithm is the need to repeat this lengthy procedure for each speaker. These disadvantages would most likely exclude this procedure from consideration in the production of commercial multi-speaker TTS systems. This procedure has instead been utilised in the production of a formant diphone-concatenation speech synthesiser which is being used in research on speech perception and voice quality and where modification of formant parameters in reasonably natural and highly intelligible synthetic speech is required.

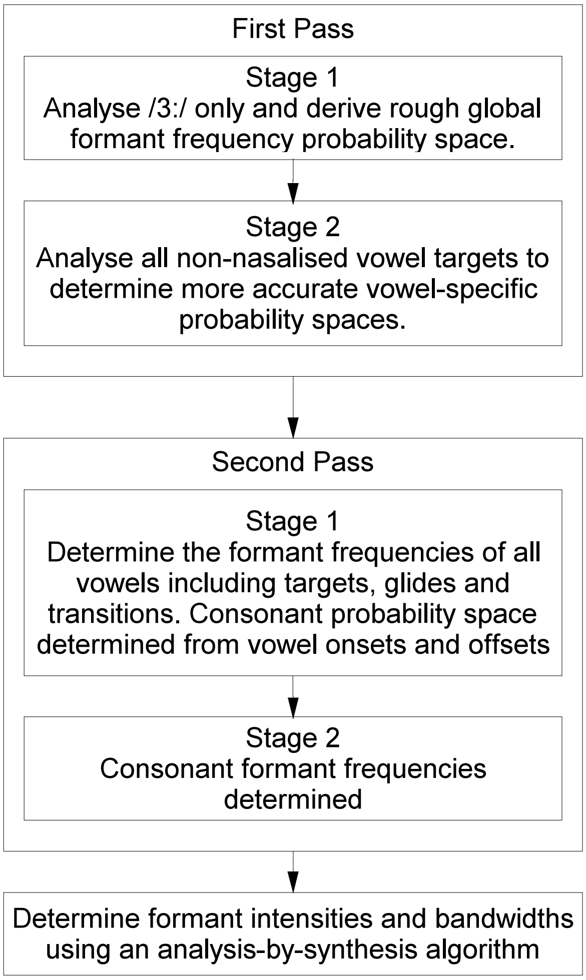
5. REFERENCES

1. Clark, J.E., Summerfield, C.D., and Mannell, R.H., "A high performance digital hardware synthesiser", *Proc. 1st Australian Conf. Speech Sc. and Tech.*, Canberra, November, 1986, 342-347.

2. Flanagan, J.L. *Speech Analysis Synthesis and Perception*, Springer-Verlag, Berlin, 1972.

3. Florentine, M., Buus, S., and Mason, C.R., "Level discrimination as a function of level for tones from 0.25 to 16 kHz", J.Acoust.Soc.Am. 81, 1987, 1528-1542.

4. Summerfield, C.D., and Clark, J.E., "Implementation of a high performance formant speech synthesiser", *Proc. 1st Australian Conf. Speech Sc. and Tech.*, Canberra, November, 1986, 354-359.



First Pass

Stage 1
Analyse /3:/ only and derive rough global formant frequency probability space.

Stage 2
Analyse all non-nasalised vowel targets to determine more accurate vowel-specific probability spaces.

Second Pass

Stage 1
Determine the formant frequencies of all vowels including targets, glides and transitions. Consonant probability space determined from vowel onsets and offsets

Stage 2
Consonant formant frequencies determined

Determine formant intensities and bandwidths using an analysis-by-synthesis algorithm

Figure 2: Overview of the formant parameter extraction procedure.