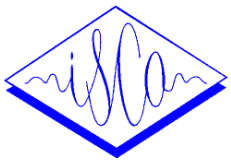

The ESCA Workshop
on Speech Synthesis
Autrans, France
September 25-28, 1990

SPECTRAL TRANSITIONS IN RULE-BASED AND DIPHONE SYNTHESIS

Douglas O'Shaughnessy

INRS-Telecommunications, Université du Québec
Nuns Island, Quebec, Canada H3E 1H6

ABSTRACT

The problem of adequate dynamic modeling of the speech spectrum is explored for general text-to-speech applications. Using analysis of formant patterns from English speech, natural formant patterns in time are compared with those produced by the MITalk system, noting where the system has difficulties in modeling spectral transitions. Phonetic contexts where a diphone approach would have the most difficulty are noted, i.e., where the diphone coarticulation assumption is invalid. To improve phoneme-based synthesis systems, better rules are needed to model coarticulation for phoneme-concatenation synthesis. To improve diphone synthesis, I enumerate contexts where triphones would better model natural speech.

1. INTRODUCTION

Most text-to-speech systems which allow general text as input concatenate small speech units (e.g., phonemes or diphones).for synthetic voice output. When concatenating phoneme units, explicit rules must be stated for modifying the stored spectral parameters, so as to simulate coarticulation effects. The most common approach stores formant frequency targets (both center frequencies and bandwidths, as well as amplitude information) for each phoneme, and adjusts these numbers to produce a time sequence of parameters every frame (e.g., every 10 msec). A large number of fairly complex rules is often needed to account for the wide variety of coarticulation effects.

A popular alternative is the diphone approach, where approximately 1000 patterns are stored corresponding to the 30 or so phonemes in a language. Obtaining and concatenating typical transitions from the actual speech of a speaker significantly reduces the number of parameter-adjustment rules needed at the synthesis stage. Some smoothing of spectral parameters (e.g., LPC coefficients) is needed at diphone boundaries, but most coarticulation effects are claimed to be accounted for within the diphones themselves, on the assumption that the phonetic effect a phoneme has on two neighboring phonemes extends only halfway through the duration of each of those phonemes. This assumption is certainly not valid for all phoneme groupings, and has not been adequately addressed in the literature. Using triphones and larger units to store coarticulatory effects over longer durations is not an attractive option in general, because of the potentially large number of units involved. Thus it is important to investigate for which phoneme contexts the diphone assumption is not valid.

2. EXAMINATION OF NATURAL SPEECH

In examining wideband spectrograms of three male speakers' voices (both for isolated words and continuous speech), I have attempted to formulate rules for improved automatic speech synthesis. The rules are based on a prototypical adult male speaker (generalized from these speakers' voices), which might well be adapted to produce different voices via simple modifications. A key point in the choice of ways to improve synthetic speech involves the spectral representation of a frame of speech. Currently, LPC or formant representations are favored for their efficiency and performance. From both speech production and perception research, we know the importance of spectral peaks (and the relative unimportance of spectral valleys and spectral tilt). While the

traditional formant representation may not be optimal (e.g., there is no real evidence that humans track formants in perception, nor concentrate on producing invariant formants in speech production), representing a frame of speech with 3 formant frequencies and bandwidth measures (for vowels) or with coarse frequency cutoff measures (for frication) is more efficient than using 10-16 LPC coefficients.

The difficulty of accurate formant tracking and the necessity of using a speech-varying representation (e.g., 3 formants for vowels, and another spectral peak measure for frication) has prevented more widespread use of formants/spectral-peaks as a speech representation. Use of a spectral peak measure does not necessarily mean that formants need be followed slavishly. When they are hard to determine (due to weak energy or merging formants), a modified representation (still spectral peak-based, but not necessarily employing one formant per kHz duly labeled as F1, F2, F3, etc) is perhaps best. Relatively few detailed rules for formant-based synthesis seem to have been published in the open literature.

A promising approach would seem to be one of triphones (i.e., phone models in context of left and right phonemes), where averaging could be done to reduce the number of models where the contextual effects of different phonemes sharing common phonetic features are very similar. For example, /l,t,d,n,s,z/ have similar coarticulation effects on adjacent vowels; thus one could use a single model for the left/right contexts of an alveolar consonant. Whether it is more efficient to cluster the triphone models based on acoustic analysis or directly based on articulatory grouping according to known phonetic features remains to be seen. An inventory of common coarticulation effects, as determined in the rules below, could help organize such a clustering.

2. A MODEL

A large majority of the relevant spectral information for speech perception lies in the "telephone bandwidth" of 300-3300 Hz. Frequencies below 300 Hz are only useful for determining voicing in obstruents The telephone bandwidth preserves the first 3 formants. Formants above F3 are essentially irrelevant or redundant for sonorant perception. Some significant information, however, is lost for identification of place of articulation for stops and fricatives. A significant acoustic feature for alveolar stop burst+frication is energy extending up to high frequency well above 3.3 kHz. Similarly, alveolar fricatives (/s,z/) are relatively strong, but have little energy below 3.3 kHz most of the time, and thus are susceptible to confusion with nonstrident fricatives under a telephone cutoff. With these exceptions, good synthetic speech should be possible using only the spectrum below 3.3 kHz.

The rules below have their basis in the movements of articulators in the vocal tract; large movements generally cause large spectral change, but in a nonlinear fashion. In general, vertical motion (raising/lowering the tongue) has a direct effect on F1. In the prototypical adult male voice (that I assume for our model), vertical motion causes up to +/-300 Hz deviation (from the norm of 500 Hz for a central vowel). Since all English consonants cause a lowering of F1 (and in roughly similar amounts), I will concentrate on the more interesting F2 and F3 behavior. F2 varies widely, from about 800 to 2200 Hz, as does F3 (from about 1700 to 3300 Hz).

For English, F3 is crucial in identifying /r/; it is the only case where F3 goes below about 1900 Hz. Combined with an F1 of about 400 Hz and an F2 of about 1200 Hz, a low F3 is a reliable indicator of /r/. (I will dispense with saying "Hz" and "about" in the rest of this paper). F2 is usually affiliated with lateral tongue position, e.g., low F2 for back vowels and high F2 for front vowels. The interesting problems come when evaluating formant behavior under coarticulation in the context of continuous speech.

2.1 Steady-state values

The average steady-state F1-F2-F3 values for one speaker's (KNS) sonorants were: /i/-300, 1950, 2600; /I/-400, 1600, 2500; /e/-400, 1850, 2500; /E/-500, 1500, 2500; /ae/550, 1500, 2400; /uh/-550, 1250, 2450; /a/-600, 1100, 2500; /O/-500, 1050, 2400; /o/-450, 950, 2000; /U/-400, 1000, 2400 /u/-300, 900, 2000; /r/-300, 1200, 1700; /l/-400, 800, 2700; /w/-300, 700, 2100.

2.2 Continuous formant transitions

Some phoneme-to-phoneme transitions involve little articulatory movement, and thus show little formant change at phoneme boundaries. Examples: a palatal fricative to or from a high front vowel /i,I,e,E/; labial /p,b,m,f,v/ or velar /k,g/ from a back round vowel /u,U,o/; alveolar fricative /s,z/ to a low or mid vowel /a,E/. The first case appears to be symmetric (i.e., little movement either to or from the consonant), whereas the latter two are not (e.g., /ug/ shows little F change, whereas /gu/ does not). Symmetry in CV and VC contexts cannot be assumed in many cases; in /ug/, the lips unround very little until after the stop closure, hence minimal coarticulation effects; but in /gu/ the lips round mostly during the /u/, hence significant formant change during the vowel.

Diphone synthesis methods often presume that spectra achieve some (perhaps brief) form of steady-state at or near the midpoint of each phone. Thus diphone units could be extracted and concatenated, yielding speech which deviates little from natural speech patterns. In practice, however, many phone sequences exhibit spectra which move continuously throughout the entire duration of individual phones. In such cases, arbitrarily choosing a segmentation point can lead to smoothing problems at synthesis time (at best) and poor spectral modeling in the output speech (at worst). It is thus worth identifying such phone sequences, so that they may be modeled as triphones (either for storage purposes, in a diphone system, or as part of formant rules, in an expert system approach).

The relatively small range of variation for F1 leads to few problems; so we confine our view to F2-3, where rapid changes of more than 400 Hz regularly occur during individual phones in certain contexts. The rules below are based on observations of words with 1-2 syllables in a brief carrier sentence (additional coarticulation occurs in more rapid speech or with longer words). In careful speech, outright undershoot of formant targets (where a formant trajectory does not achieve a value normally attained in most other contexts) is not as common as might be expected, but does occur reegularly in F2 for alveolar+back rounded vowel+alveolar sequences and in /l/+high front vowel+/l/. Much more common are cases where a formant starts rising/falling during one phone and continues that movement smoothly throughout the ensuing phone without any steady state: F2 for liquid+high front vowel+velar, palatal fricative+/E/+labial fricative, alveolar+/u/+labial, alveolar+/Es/; F3 for alveolar+/E/+velar, /rE/+alveolar, /jUr/, /roz/, /wEl/, and others.

In comparing natural F2-3 patterns in these cases against those predicted by MITalk, I found that the transitions tended to be well modeled in terms of the initial and final values, but not necessarily in the dynamics of the transitions. MITalk tended to model many large formant movements with patterns of slow change at phoneme boundaries and rapid change in the middle of the phoneme. The natural patterns tended instead to be more of a continuous (often linear) rise or fall, rather than concentrating most of the change in the middle of the phoneme. For example, in "duke," MITalk predicted a rapid 600 Hz fall in F2 (from 1600 to 1000), followed by an equally-long period of flat F2, whereas the actual pattern showed a linear F2 fall. In the absence of perceptual tests, it is not clear how important such timing differences are to the quality of the resulting synthetic speech. In any event, the fact that these common phoneme sequences lead to rapid formant transitions without a steady state in the middle of the phoneme leads to significant difficulty for diphone modelling. The normal steady values noted above are usually attained at some point in the formant movements, but their timing is very variable. Simple diphone segmentation and concatenation will not work for these cases. For example, in "deuce," the low F2 vowel target is only attained late in the vowel, whereas in "desk" the F2 value typical of /E/ occurs very early in the vowel.

2.3 Low vowel context

Larger coarticulation effects generally occur with larger tongue movements. With low vowels /a,ae/ adjacent to consonants (which all require a relatively high tongue position), F1 exhibits significant swings of +/- 400, typically in transitions of 50 ms. The F2-3 changes in such cases are more variable. In the CV (consonant+vowel) case, velars and palatals obey a "locus" of 2000, with F2 falling from 1750 to 1250 and F3 rising from 2250 to 2500. The VC velar case is spectrally symmetric in the amount of formant change, but the timing varies: a slow 70 ms for the opening CV, a faster 40 ms for the closing VC, the latter involving a more ballistic (and hence more rapid) motion. For alveolar closures /t,d,n/ and low vowels, the formant changes are smaller: for the CV, F2 falls from 1400 to 1200 in 50 ms (and no change in F3). (In this discussion, unless otherwise explicitly stated, I assume symmetry in CV and VC patterns, but present the results only for the CV direction, for simplicity.) For labials /p,f,m/ and low vowels, the formant changes are also small: for the CV, F3 rises only 200 (and no change in F2).

2.4 High front vowel context

The coarticulation between a consonant and a high front vowel is entirely in lateral articulation; thus relatively little F1 change occurs, but large lateral motion can heavily affect F2-3. In the voiced VC case, velar /ig/ obeys a (raised) locus at 2200, with F2 rising from 1800 to 2100 and F3 falling from 2500 to 2300, in 30 ms. For the unvoiced VC (/Ik/), the F2-3 changes (in the same directions) are only 100 Hz. Furthermore, for the CV case (/ki,gi/), little F2-3 motion is seen (ditto for palatal fricatives). For alveolars and dentals, the CV case exhibits larger change: an F2 rise from 1700 to 2100 and an F3 rise from 2500 to 2800, in 45 ms; in the VC case, F2 falls from 2050 to 1850 (before voicing ceases), with little change in F3. For labials /p,b,f,v,m/ (again using the CV case), F2 rises from 1750 to 2000 and F3 rises from 2300 to 2600; the rise is slow (70 ms) in the CV case (except for nasals), and (the fall) shorter (30 ms) in the VC case.

2.5 High back vowel context

Some of the biggest coarticulation effects are found between a non-labial consonant (e.g., /t,d,k,g/) and a high back vowel. This is due to the combination of lateral articulation and lip rounding. In the CV case, velar /gu/ obeys a (lowered) locus at 1800, with F2 falling from 1500 to 1000 and F3 rising from 2000 to 2300, in 80 ms. Also in the CV case, labial consonants have a slight F2 fall from 1200 to 1000 and an F3 rise from 2100 to 2300, in 50 ms (the VC /uv/ also has this behavior - reversed in time, of course). In /wu/, a small F2 rise of 100 is seen. For the (other) VC cases, both velar and labial consonants exhibit little F2-F3 motion. For alveolars (using the CV case), F2 falls from 1750 to 1150 and F3 also falls, from 2500 to 2100; the F2 fall is slow (80 ms) in the CV case and (the rise) shorter (50 ms) in the VC case. In the CV case of an unvoiced alveolar stop (e.g., /tu/), the F2 fall occurs after the VOT (and not during the aspiration period); in the VC case of /n/, the changes are smaller (i.e., F2 only rises to 1200 at the /n/ onset). Similarly, fricatives (/su,uS/) shorten the F2 change: 1500 at the boundary. Dental fricative effects are similar to those of alveolars, but smaller: e.g., a 1300 F2 boundary.

ACKNOWLEDGMENT: This research was supported by the Quebec FCAR program.

REFERENCES

1) Allen J. (1976) "Synthesis of Speech from Unrestricted Text," Proc. IEEE, vol. 64, 433-442.
2) Dettweiler H. & Hess W. (1985) "Concatenation rules for demisyllable speech synthesis," Acustica, vol. 57, 268-283.
3) Klatt D. (1980) " Software for a cascade/parallel formant synthesizer," J. Acoust. Soc. Am., vol. 67, 971-995.
4) Shadle C. & Atal B. (1979) "Speech synthesis by linear interpolation of spectral parameters between dyad boundaries," J. Acoust. Soc. Am., vol. 66, 1325-1332.