FROM SEGMENTAL SYNTHESIS TO ACOUSTIC RULES USING TEMPORAL DECOMPOSITION.

Frédéric BIMBOT, Gunnar AHLBOM, Gérard CHOLLET

ENST Dept. SYC, CNRS UA-820
46 rue Barrault, 75634 PARIS cédex 13, FRANCE.

ABSTRACT

A methodology is proposed to infer automatically acoustic rules that could be used to predict natural spectral transitions for speech synthesis. It adapts ATAL's "temporal decomposition" technique /1/ to compute interpolation functions from phonetically labelled acoustic targets. Coarticulation effects are controlled quite adequately using such a representation. With this methodology, rule-based synthesis will be developped more efficiently for new languages, dialects, speakers with better control of speaking rate, style of speech ...

INTRODUCTION

The automatic generation of "natural" speech from a phonetic transcription is a challenging task. Two main approaches have been proposed: segmental and rule-based. The **segmental** approach (using diphones, demi-syllables, polysons, ...) offers an easy way to intelligible speech. But the segment inventory is speaker dependent and control of timing is a non trivial task.The lack of naturalness could be attributed to uneasy analytic control of speech parameters. A **rule-based** approach is more flexible, gives more insight on the perceptually relevant features of speech, and could be more easily adapted to new speakers. Control of prosody, style of speech, is achieved quite naturally within a unified framework. Unfortunately, this approach requires, so far, a lengthy and art oriented procedure using visual and auditory hand-tuning of the rules.

Our goal is to provide a methodology to move gradually from segmental to rule-based approaches. We propose a number of interactive tools using powerful signal and data analysis techniques to model spectral evolution, infer spectral targets automatically, and generate adequate transitions toward these targets.

SYNTHESIS and COARTICULATION

An acoustic synthesizer is usually controlled by a set of parameters updated at regular time intervals. The parameters are either retrieved from memory (speech restitution and segmental synthesis) or computed from rules. We are concerned here with smooth spectral evolution corresponding to articulatory dynamics. As an working hypothesis, articulatory and therefore spectral targets are assumed. In this paper, coarticulation is referred to as a phenomenon of target undershoot due to contextual effects, speaking rate ...

TEMPORAL DECOMPOSITION

ATAL's technique /1/ decomposes speech into phone-length temporal events which could be interpreted as overlapping and interacting articulatory gestures /2, 3, 4/. Evolution of a sequence of m spectral vectors $[y_i(n)]$ is approximated as a linear combination of m events represented by known functions $\varnothing_k(n)$ (interpolation functions) with appropriate weights y_{ik}(targets):

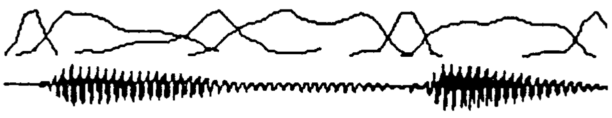
$$y_i(n) = \sum_{k=1}^{m} y_{ik} \, \varnothing_k(n)$$

The functions $\varnothing_k(n)$ are constrained to be compact in time: that is zero everywhere except on a segment. The first step of the algorithm consists in finding a good approximation for the localization and the extent of the $\varnothing$-functions. Once a set $\{\varnothing_k\}$ has been found, the corresponding target vectors y_k are computed by:

$$[y_k] = [y_{ik}] \, [\varnothing_k]^t \, ([\varnothing_k] \, [\varnothing_k]^{-1})^t$$

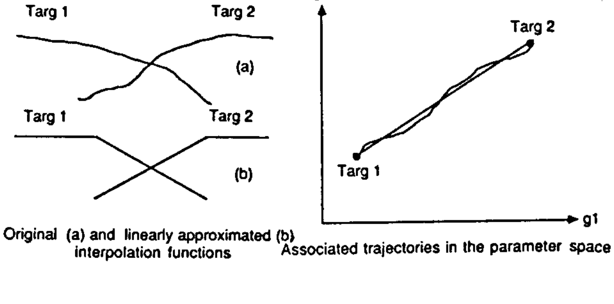
which minimizes the reconstruction error according to a least square criterion.

Iterative refinement can then be performed until no significant improvement is obtained.



Temporal decomposition of the speech segment /ede/.

$\varnothing$-functions can be linearly approximated and normalized so that their sum be constant and equal to unity. With this approximation, temporal decomposition of a speech segment correspond to a piece-wise linear trajectory in the parameter space.



Original (a) and linearly approximated (b) interpolation functions

Associated trajectories in the parameter space

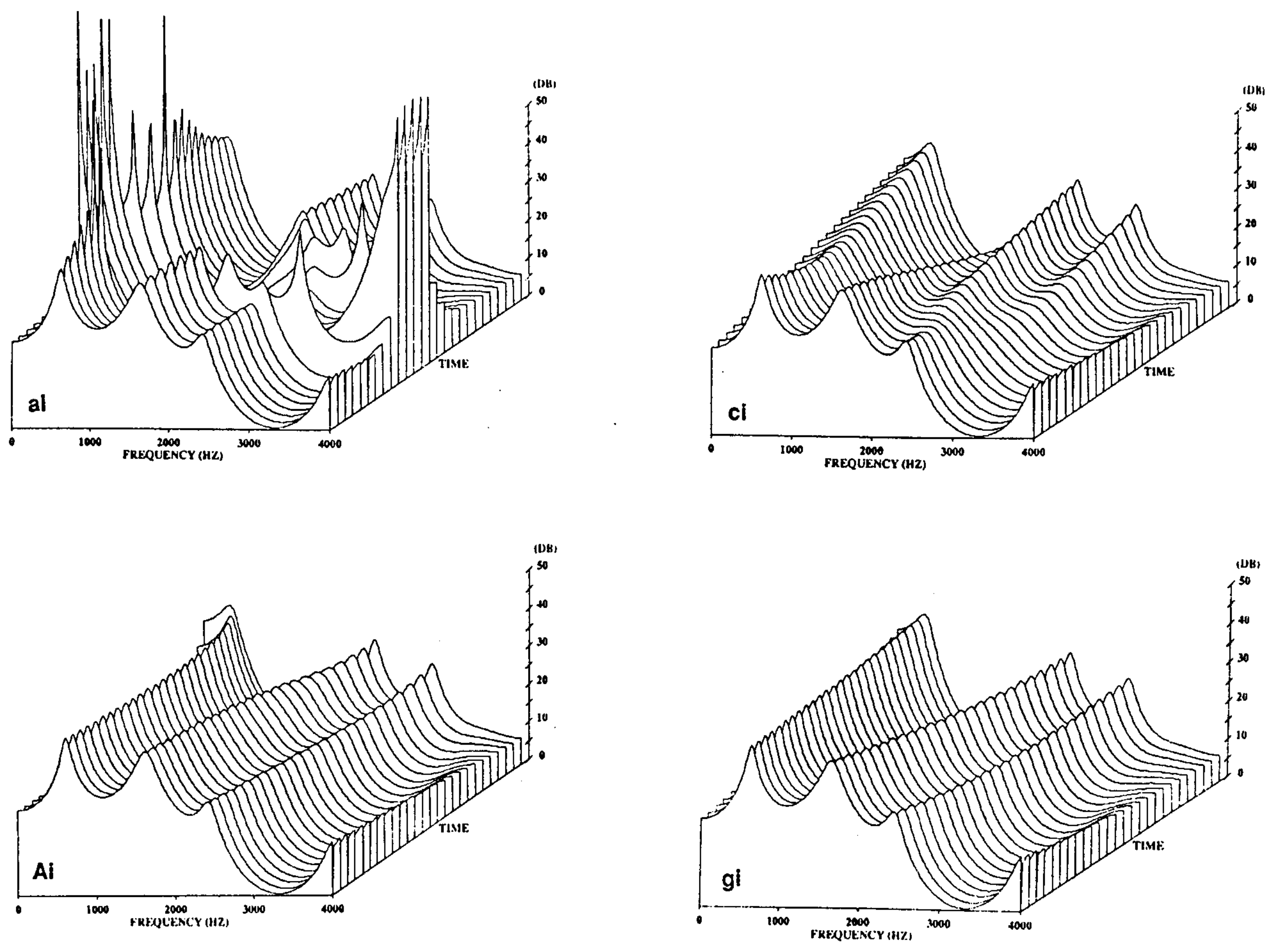


Fig. I Synthetic spectrums associated to a linear trajectory between 2 targets, in different spectral spaces:
ai: auto-regressive coeff, ci: cepstral coeff,
Ai: area parameters, gi: log area ratios.

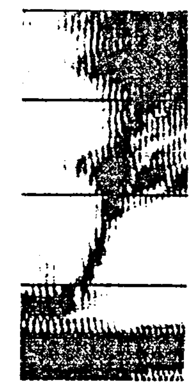
SPECTRAL REPRESENTATIONS

A description of transitions is attempted as a linear combination of spectral parameters. A number of spectral representations could be used for this purpose /5/:

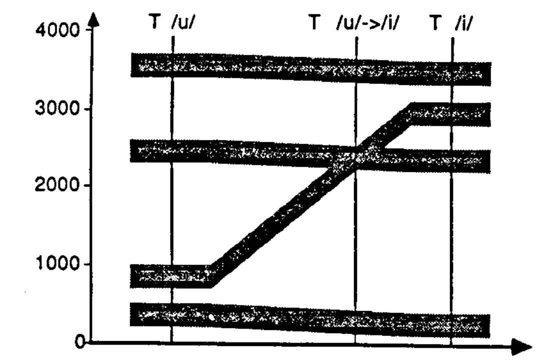
Formant frequencies, amplitudes, and bandwidths (F_i, A_i, BW_i) are often used for speech parameterisation, owing to their physical meaning. However, they necessitate a labelling operation. Moreover, a complex treatment must be performed in order to interpolate spectra with different number of formants. Poles (z_i) and line spectrum pairs (LSP_i) have the same drawbacks.

We therefore investigated the effect of interpolating spectral parameters for several unlabelled spectral representations: LPC autoregressive coefficients (a_i), cepstral coefficients (c_i), area parameters (A_i), reflexion coefficients (k_i), and log area ratios (g_i) /6/.

Auto-regressive coefficients are inadequate as the associated space is not linearly stable. Cepstral coefficients are neither suitable since the mean of two vectors (c_i) gives a spectrum which keeps the peaks of both original spectra. Area parameters seems more convenient, but the interpolated formant trajectories are not quite linear. Reflexion coefficients (k_i) behave adequately with damped resonances. Log area ratios (g_i) are the best parameters we have found so far (see fig. I).



Spectrogram of speech segment /ui/



Typical transition between phonemes /u/ and /i/

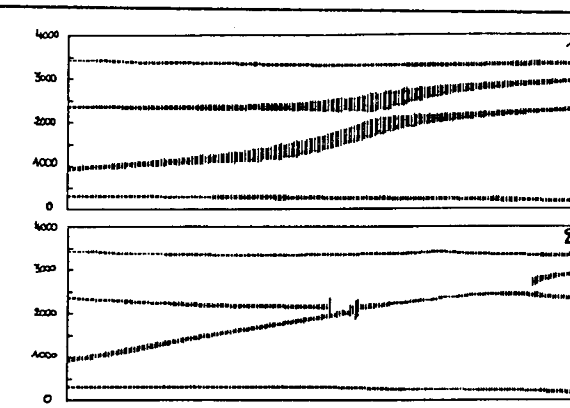
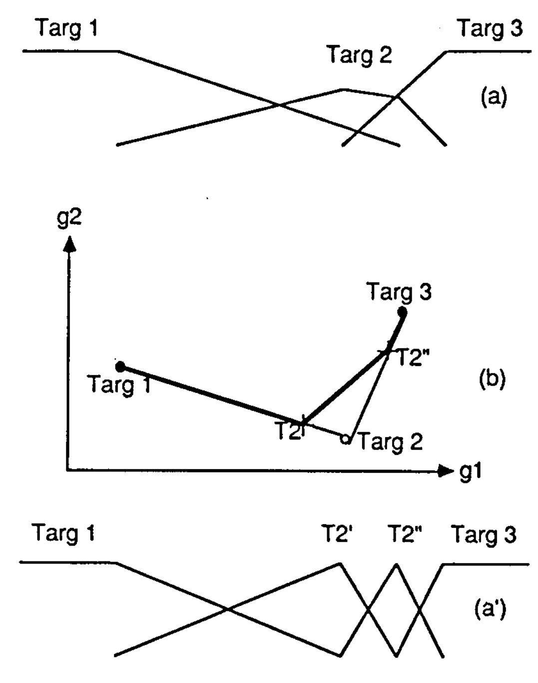


Fig. II Synthetic spectrums obtained by a linear interpolation between log area ratios of typical /u/ and /i/ targets.
(1) without intervention of any intermediate target,
(2) with adjunction of an adequate intermediate target.

COARTICULATION

An analysis of temporal decomposition results reveals the acoustic-phonetic structure of speech. Quasi-stationary segments (fricative, nasal consonant, vowel nuclei) are described with a single function. Transitions are usually described with two overlapping Ø-functions /7/. But some transitions require an extra function $Ø_2$ associated with Targ 2:



Temporal patterns (a) and (a') give different descriptions of the same trajectory (b).
(a) is a best description of the actual articulatory gesture (undershot target)

$Ø_2$ usually describe a highly coarticulated phone with undershoot of the corresponding target. In other cases, the extra function Ø is a correction of the trajectory between targets 1 and 3. This is the case for a rapid front-back movement of the tongue (in such diphones as [ui], [wi], [iu], and [ju]), which correspond to a "crossing formant" configuration /8/. The existence of an intermediate target renders more accurately the spectral transition (see fig. II).

SEGMENTAL SYNTHESIS

Synthesis can be achieved successfully by concatenating stored segments . A set of such segments called "polysons" is chosen in such a way that coarticulation effects accross boundaries are minimized /9/. This is achieved by placing boundaries on spectrally stable sounds (vowels, fricatives, nasal consonants, occlusion of plosives). About 7000 "polysons" were selected for French synthesis. Significant improvement in perceptual quality (intelligibility and naturalness) is achieved with "polysons" synthesis as compared to diphone synthesis. Unfortunately the number of these units is an order of magnitude larger than the number of diphones.

Temporal decomposition can be used to encode "polysons" very efficiently /10/.

RULE-BASED SYNTHESIS

"Polysons" are being classified according to the structure of their Ø-functions /10/. For instance, the temporal patterns of all combinations of a vowel and an unvoiced fricative (/as/, /if/, /us/) are similar.

The archetype of each group can be viewed as a rule to synthesize "polysons" of that group. A "polyson" is therefore reduced to a Ø-pattern type and a set of associated targets.

The edges of "polysons" are quasi-stationary segments, described with a single Ø normalized to unity. The concatenation of "polysons" is restricted to those with matching targets on edges (much like dominos).

The Ø-pattern can be distorted by rules to take care of variations in speaking rate, stress, emphasis... Overlapping and smoothing of Ø-functions at boundaries express the coarticulation effects accross "polysons".

CONCLUSIONS

Temporal decomposition using target spectra can break the complex encoding of these segments. In particular, coarticulation effects are analyticaly explained and modeled. It is demonstrated that these new tools provide an adequate environment in our search for better rules in acoustic speech synthesis.

ACKNOWLEDGMENTS

Some ideas developped in this work where discussed with colleagues from IPO, Eindhoven, during a sabbatical year G. CHOLLET spent there. Contributions of S.M..MARCUS were particularly important in the initial phase.

REFERENCES

/1/ ATAL B.S. **Efficient coding of LPC parameters by temporal decomposition.** Proc. ICASSP-83, 2.6, 81-84, 1983.

/2/ MARCUS S.M., Van LIESHOUT R.A.J.M. **Temporal decomposition of speech.** IPO annual progress report 19, p. 25-31, 1984.

/3/ AHLBOM G., F. BIMBOT, G. CHOLLET. **Modeling spectral speech transitions using temporal decomposition techniques.** ICASSP, Dallas, 1987.

/4/ CHOLLET G., GRENIER Y., MARCUS S.M. **Temporal decomposition and non-stationary modeling of speech.** EUSIPCO, The Hague, 1986.

/5/ SCHAFER R.W., RABINER L.R. **Parametric representations of speech.** From: Speech Recognition, REDDY R. (ed.), 1975.

/6/ VISWANATHAN R., MAKHOUL J. **Quantization properties of transmission parameters in linear predictive systems.** IEEE Trans. ASSP 23, pp. 309-321, 1975.

/7/ OHMAN S.E.G. **Coarticulation in VCV utterances: Spectrographic measurements.** JASA 39, pp. 151-168, 1966.

/8/ CHAFCOULOFF M., CHOLLET G., DURAND P., GUIZOL J., RODET X. **Observation and modeling of the variability of formant transitions using ISASS.** IEEE-ICASSP, Denver, 1980.

/9/ CHOLLET G., GALLIANO J. F., LEFEVRE J. P., VIARA E. **On the generation and use of a segment dictionnary for speech coding, synthesis and recognition.** IEEE-ICASSP, Boston, 1983.

/10/ BIMBOT F., CHOLLET G., MARCUS S. M. **Localisation et representation temporelle d'evenements phonetiques: applications en etiquetage, segmentation et synthese.** JEP-86, Aix-en provence, 1986.