FROM SEGMENTAL SYNTHESIS TO ACOUSTIC RULES USING TIME DEPENDENT MODELING TECHNIQUES.

Gérard CHOLLET*, Gunnar AHLBOM*, Frédéric BIMBOT*, Alvaro DE LIMA-VEIGA*.

ABSTRACT

Intelligible Text-to-Speech may be achieved by concatenating spectrally encoded segments. However, its lack of naturalness could be attributed to a difficult control of speech parameters. Acoustic rules are more adequate for this control. The aim of this work is to provide a methodology to move from a segmental to a rule-based approach. A number of interactive tools is proposed using powerful signal and data analysis techniques for modeling spectral evolution, inferring spectral targets, and generating adequate transitions between these targets. The choice of adequate spectral parameters is essential. A set of French speech segments ("polysons") of a single speaker has been encoded using these tools. Spectral targets were constrained to belong to a finite set of vectors (allophonic targets). Coarticulation effects (vowel reduction, nasalisation...) can be accounted for by controlling the time duration of temporal evolution functions. Segment concatenation problems are eliminated. Automatic procedures to select allophonic targets for new speakers and group temporal patterns into rules are the current issues.

UNLIMITED VOCABULARY SPEECH SYNTHESIS

Two main approaches have been proposed for unlimited vocabulary speech synthesis. 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. Its lack of naturalness can be attributed to uneasy analytic control of speech parameters. A rule-based approach is more flexible, gives more insight on the relevant features of speech, and may allow speaker modification. Control of prosody, style of speech, sentence rhythm, is achieved quite naturally within a unified framework. Unfortunately, this approach requires human knowledge and manual intervention, for visual and auditory hand-tuning of the rules. The time-dependent modeling techniques, described below, permit a structural description of spectral evolution in speech segments. From their results, rules can then be obtained by inferring spectral targets and extracting typical transitions patterns between these targets.

MODELING TEMPORAL EVOLUTION

Speech can be encoded using a sequence of p-dimensional "spectral" vectors $y(t)$, corresponding to a time sampling of the vocal tract transfer function. A synthesizer could, either retrieve from memory an appropriate sequence, or generate this sequence by rules. The rules should predict the vector $y(t)$ for every time instant $t$. In order to infer these rules, two analysis techniques have been experimented with: vectorial AR models and temporal decomposition.

Vectorial AR models

De Lima-Veiga and Grenier /Ref.1/ propose to model the evolution of spectral vectors $y(t)$ as a vectorial AR process driven by a piece-wise constant function v(t):

$$y(t) = \sum_{i=1}^{p} A_i y(t-i) + v(t)$$

The driving vector $v(t)$ may represent the current spectral target, and hence be expressed as:

$$v(t) = v(t-1) + u(t)$$

where $u(t)$ is a zero-valued function almost everywhere except at instants of changes of the spectral target $v(t)$. Identification of the model is made by applying classical vector AR

* CNRS UA-820, ENST, 46 rue Barrault, 75634 PARIS Cédex 13, FRANCE
where \( u(t) \) is a zero-valued function almost everywhere except at instants of changes of the spectral target \( v(t) \). Identification of the model is made by applying classical vector AR estimation methods, which gives the matrices \( A_1 \ldots A_n \). Jump amplitudes and times are estimated in order to minimize a least square reconstruction error. Jumps are placed one at a time until a predetermined error improvement is achieved. Perfect reconstruction is to be avoided as this may imply a number of jumps being too large (Fig.1). The entry \( v(t) \) contains the informations on the values of successive spectral targets, and the timing of jumps from one to another. The matrices \( A_j \) describe the local transient behavior of the spectral parameters, for instance time-constants of the vector filter.

**Temporal decomposition**

ATAL's technique /Ref.2/ decomposes speech into phone-length temporal events, which could be interpreted as overlapping and interacting articulatory gestures /Ref.3,4/. The evolution of a sequence of spectral vectors \( y(t) \) is approximated as a linear combination of \( n \) events, represented by known functions \( \phi_k(t) \) (interpolation functions) with appropriate weights \( y_k \) (targets):

\[
y(t) = \sum_{k=1}^{n} \phi_k(t) y_k
\]

The functions \( \phi_k(t) \) 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 \( \phi \)-functions. Once a set \( \{ \phi_k \} \) has been found, the corresponding target vectors \( y_k \) are computed by minimizing the quadratic reconstruction error. Iterations on the computation of \( \phi_k \) and \( y_k \) can then be applied to improve the results or to fulfil extra constraints (Fig.2).

![Comparison of speech segment and reconstructed LAR-coefficients](image)

*Fig. 1 shows the speech segment /alu/ and below the original and reconstructed LAR-coefficients by using Vector AR modeling.*

*Fig. 2 represents the same thing but obtained by using temporal decomposition.*

**PARAMETRIC REPRESENTATION**

\( \phi \)-functions can be normalized so that their sum be constant and equal to unity. With this approximation, the evolution of the spectral vectors corresponds to a piece-wise linear trajectory in the parameter space. Such a property is obtained on the vector AR model results by constraining all time-constants to be equal. The nature of spectral parameters is therefore crucial /Ref.5/. Linear combinations of them should rely on an acoustic signification. Previous studies showed that Log Area Ratios /Ref.6/ were suitable parameters for spectral interpolation: an arithmetic progression between two LAR-vectors correspond to a linear trajectory of the associated formants /Ref. 7/ (Fig.3).

The temporal decomposition technique can produce directly this result if linearly approximated \( \phi \)-functions are used in the LAR space. A similar configuration is obtained with the AR-vectorial technique, if models operate on Area Ratios instead of Log Area Ratios.
EXPERIMENTS and OBSERVATIONS

Temporal decomposition describes quasi-stationary segments (fricative, nasals, vowels) with a single function. Transitions usually necessitate two overlapping $\emptyset$-functions, but some of them require an intermediate extra function that is a correction of the trajectory between extreme targets (Fig. 4). This is the case for rapid front-back movements of the tongue (in such diphones as [ui], [io], [ju]), which correspond to a "crossing formant" configuration /Ref. 8/. The existence of an intermediate target renders more accurately the spectral transition. Highly coarticulated phones, in context of two steady regions, show the same $\emptyset$-pattern, that express the undershoot of the corresponding target.

With vector AR modeling techniques, the same phenomenon is described by a jump towards a new target before the previous one is reached.

TOWARD RULE-BASED SYNTHESIS

Defined as a segment of speech bordered by two spectrally stable sounds (vowels, fricatives, nasals, plosive occlusion), the "polyson" was first studied as a suitable unit for segmental synthesis. Coarticulation effects at the boundary between polysons are indeed minimized. Experiments in French showed significant improvements in perceptual quality with polyson synthesis, as compared to diphone synthesis. However, needed storage is about ten times larger.

Models of spectral evolution of polysons have been studied in order to describe their structure and infer synthesis rules automatically.
Allophonic spectral target inference

Whatever the context of a given stable phoneme, all associated targets are neighbours in the spectral space; clustering them allow the extraction of labeled reference targets for frontier phonemes of polysons. Once this primary set is obtained, original border targets are replaced by their representant, and intermediate targets are re-evaluated to insure coherence. Vector quantization is then performed on the secondary collection of targets, and allophonic reference targets for unstable phonemes are obtained and labeled (voiced / devoiced [j], vocalic / consonantic [R], front / central / back vocalic [i], [iu] trajectory correction, ...).

Time dependent pattern clustering

Polysons are classified according to the structure of their temporal evolution. Diphonic polysons generally give simple temporal patterns that are easy to group. For instance, the temporal patterns of all combinations of a vowel and an unvoiced fricative ([as], [if], [us]) are similar. More complex is the extraction of typical patterns for such polysons as [ilja], [ilje], [iljo]... [alja]... [ulja]...

Defining acoustic rules

The archetype of each group can be viewed as a rule to synthesize polysons of that group. A polyson is therefore reduced to a temporal pattern type and a set of associated targets. For instance, spectral parameters of [kRi] and [tRu] are regenerated with the same temporal pattern (as members of the class: "unvoiced plosive + [R] + vowel"), but with different spectral targets ([k]tar / [t]tar, [consonantic R]tar, [i]targ / [u]targ).

The temporal pattern can be changed by rules to take care of local prosodic variations (stress, emphasis, vowel falls), as well as global supra-segmental phenomena (speaking rate, articulation characteristics)...

Substituting sets of reference targets with one another may allow for multi-speaker synthesis.

CONCLUSION

Time dependent modeling techniques, using target spectrums inference and temporal evolution description, can break the complex encoding of speech segments. In particular, coarticulation effects are analytically explained and modeled. Thanks to these new tools, rules for synthesis can be expressed and extracted from the observation of results obtained on polysons.

ACKNOWLEDGMENTS

Two students from ENST contributed to software developments for these studies. A. VIGIER contributed to the PASCAL implementation of the temporal decomposition techniques, and P. KERHARO developed some programs for target inference. We would like to thank them and acknowledge their contribution.

REFERENCES

[Ref.1/ DE LIMA-VEIGA A., GRENIER Y. Two level Bernoulli-Gaussian modeling of non-stationary signals. (in preparation)
[Ref.2/ ATAL B.S. Efficient coding of LPC parameters by temporal decomposition. Proc. ICASSP-83, 2.6, 81-84, 1983.