

SPEECH SYNTHESIS BY STRUCTURED SEGMENTS, USING TEMPORAL DECOMPOSITION AND A GLOTTAL EXCITATION

F. BIMBOT, G. CHOLLET, P. DELEGLISE.

E.N.S.T. - Dépt SIGNAL, C.N.R.S. - URA 820, 46 rue Barrault, 75634 PARIS cedex 13, FRANCE, Europe.

Abstract :

Classical speech synthesis systems either concatenate diphone-like tabulated patterns or reconstruct speech parameters according to pre-defined rules. Both techniques show drawbacks : the former lacks flexibility while the latter is highly time-consuming to build.

We propose an intermediate technique using structured segments : segmental units are still resorted to, but they are automatically analysed in terms of a set of spectral targets, a temporal decomposition pattern and a parametric glottal excitation.

Structured segments can then be handled by rules. They also supply a valuable material which can be referred to, for building gradually a reconstructive synthesis system.

Introduction :

Linguists describe speech as a sequence of elementary sounds, called phones. Indeed, a phonetician can transcribe any sentence into a string of consecutive units, arising from a limited set of symbols. However, the observation of a speech signal reveals a continuum of acoustic properties; borders between successive phones can not be sharply decided. Moreover, there does not exist any fixed spectral characteristics for each phone : to a unique phonetic symbol corresponds an infinite set of realizations.

Fant resolves this apparent paradox in explaining that the phonetic string describes the informative content of a message, while the speech signal is the result of the coding of this message by the human speech production system [1]. From the mechanical inertia and the motorial inaccuracy of human articulators, result coarticulation and variability phenomena.

The task of speech recognition consists in decoding the speech code, as Marcus mentioned it [2]. On the opposite, speech synthesis aims to encode a phonetic string into an acoustic signal, which will be easily made out by the listener, and will if possible sound natural to him. In such a framework, coarticulation effects and variability consequences have to be rendered.

Rule-based speech synthesis systems reconstruct a set of control parameters according to pre-defined rules. Designed to be flexible, such a system is however very time-consuming to build, for it needs an empirical ear-tuning of the rules.

The segmental approach can therefore be considered as an expedient to pass round this major drawback. In segmental systems, control parameters are obtained by concatenating pre-encoded units, which are classically diphones. Although it is an easy way to synthetic speech, the lack of structuration of these tabulated patterns make them very difficult to handle for prosodic modifications.

We experimented a new synthesis technique, which is intermediate between segmental and rule-based approaches. Segmental units are still resorted to, but they are automatically structured by a temporal decomposition analysis. The structured

segments can then be handled by rules, for prosodic modifications. A glottal excitation is used to increase speech quality and naturalness.

A methodology is finally proposed for moving gradually towards a rule-based speech synthesis system, by building automatically a set of acoustic knowledge on speech. Possible enrichments of the temporal decomposition technique and alternative spectral evolution modelings are also discussed, as a conclusion.

Temporal decomposition :

Atal proposed a technique of temporal decomposition of speech, for coding purposes [3]. The evolution of N spectral parameters (m -dimensional vectors) $y(t)$ is approximated as a linear combination of a limited number n of m -dimensional targets vectors g_k , weighted by scalar compact interpolation functions $\phi_k(t)$:

$$y^{\#}(t) = \sum_{k=1}^{k=n} g_k \phi_k(t), \quad \text{i.e. } Y^{\#} = G \phi,$$

where $y^{\#}(t)$ denotes an approximation of $y(t)$.

ϕ -functions express the contribution of the targets over the speech segment. They are constrained to be non-zero on a segment only (compacity), so that the influence of a target is limited in time. Speech is thus described as temporally overlapping events of limited duration.

As a first step, each ϕ -function is searched, as the best linear combination of the first principal components of original spectral vectors (obtained from a singular value decomposition), according to a compactness criterion. Targets are then computed by minimizing the mean square reconstruction error of the original vectors. Iterative refinement of both interpolation functions and spectral targets is finally done. Atal uses LPC-derived log area parameters.

We developed some extensions to Atal's method, in order to increase the phonetic relevance of the technique. The robust temporal decomposition algorithm resorts to an iterative estimation of ϕ -functions during the first step, using local singular value decompositions and an adaptive windowing [4]. We also express the compacity constraint with a slightly different formulation. We use log area ratios. Moreover, we recently experienced that a removal of the continuous component from original vectors before computing SVD makes it easier to decide the number of principal components that are to be kept for estimating ϕ -functions; provided this bias is re-introduced for the estimation itself.

The robust temporal decomposition technique ensures a higher accuracy in the description of coarticulated phones, an improved decorrelation of similar neighbouring sounds and an increased detection of plosive bursts.

Figure I gives an example of a robust temporal decomposition.

Speech portions described by a single interpolation function correspond to time interval of quasi-stability. Non-stationary segments are showed up when \emptyset -functions overlap. Local sharp curvatures of the spectral trajectory manifest themselves through more than 2 overlapping functions. Such configurations may correspond to highly coarticulated phones (undershot target). They may also result from some special acoustic phenomenon (formant crossing, for example) or from some asynchrony in articulation (nasalization delay, for instance) : a specific function is then needed to model these contextual events. Nevertheless, in the last case, the specific functions may also intervene alone.

Temporal decomposition gives a representation of speech in terms of time limited overlapping events, which are characterized by both a spectral target and a temporal influence. Modeling the acoustic manifestations of coarticulation contributes in undoing its effects, which is a desirable pre-processing for recognition tasks. For synthesis purposes, temporal decomposition analysis provides a flexible description of the acoustic structure of speech, though still allowing a straightforward reconstruction.

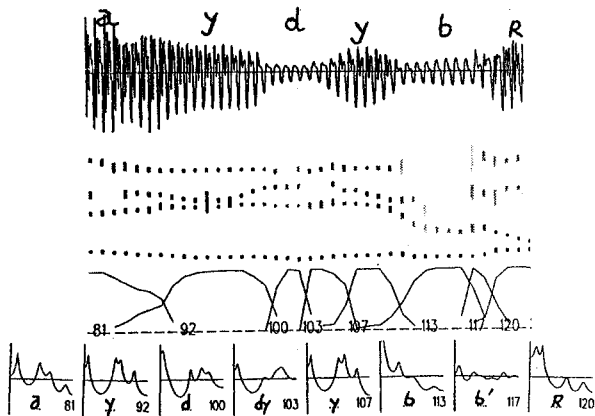


Figure I : Temporal decomposition of speech segment [aydybr].
phonetic transcription, schematic LPC spectrum,
interpolation functions and spectral targets.

Glottal excitation :

A very classical model of speech production consists of an excitation waveform inputting a time-varying linear filter. This filter would approximate the transfer function of the vocal tract, whereas the excitation would represent the sound wave at the larynx. When estimating this simplified model, hypothesis have to be made on the source structure, and a minimisation criterion has to be expressed.

Voiced sounds are produced by regular vibrations of the vocal folds. The acoustic result is a pseudo-periodical speech waveform. Unvoiced sounds are obtained by a random excitation, the passing of which through the vocal tract creates a colored noise. Mixed excitations may occur.

Classical LPC modeling of voiced speech assumes a unique pulse as an excitation for a fundamental period. Multi-pulse approaches allow more than one. Still, these techniques fail in deconvolving the source from the vocal tract, for the source models are not suited to the speech production mechanisms.

Further physiological insight of the vocal folds functioning shows that an entire excitation cycle can be subdivided into 3 phases :

- 1 - the closed glottis phase while no air passes through the larynx,
- 2 - the glottal opening phase during which the vocal folds move aside from each other,
- 3 - the glottal closing phase at the end of which the air flow is stopped by their gathering again (glottal closure).

A typical glottal waveform is depicted on figure II. Fant, for instance, suggested a possible mathematical model for it [5].

Hedelin proposed an analysis technique for extracting automatically a glottal excitation from natural speech [6]. Under this approach, a glottal cycle is fully described by a few parameters : an absolute duration T , an amplitude A , a relative opening duration t_o and a relative closing duration t_c (cf. figure II). The closed phase is modeled by a constant zero function, the glottal opening by half a period of a square sine and the glottal closing by a quarter period of cosine. Phases of these functions are easily adjusted to ensure continuity at the junctions. Their amplitudes are necessarily equal to A .

The algorithm then consists in estimating at the same time, and for each glottal pulse, a position, a parametric shape and an associated all-pole vocal tract filter. Since this is a non-linear problem, an iterative procedure is used. An important computation power is required (100 times more than classical LPC), as well as a high quality of recording devices (free of non-linear phase distortion).

Speech reconstruction starts by the regeneration of the glottal pulses from their parametric representation. The resulting waveform is then used as the excitation for the successive filters, as in classical LPC.

The quality of the reconstructed speech is excellent; differences between original and synthetic speech are hardly noticeable. Moreover, realistic modifications of glottal parameters can be supported without altering significantly the subjective quality of the resulting speech; in particular, neither buzziness nor dullness arise. A high quality control of intonation, accentuation and even voice texture is made thus immediately possible. This property is of major interest for speech synthesis purposes.

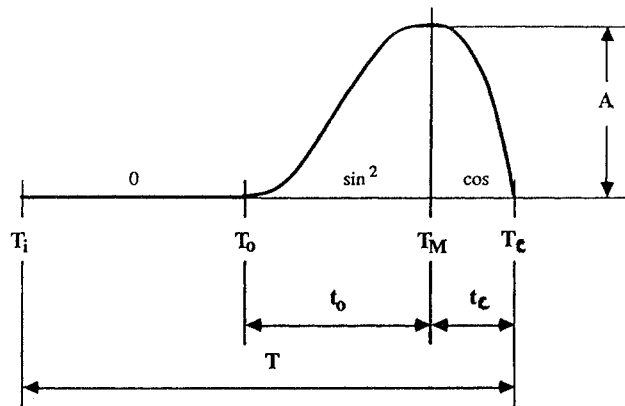


Figure II : The glottal model used by Hedelin.
 T_o , T_o' , T_M and T_c : absolute instants.
 T , A , t_o , t_c : glottal parameters.

Speech synthesis by structured segments :

Whereas rule-based speech synthesis systems reconstruct control parameters from human-tuned rules, segmental synthesis resorts to concatenations of tabulated pre-encoded units; a kind of montage. If flexible, the first approach requires a considerable work for adjusting the rules, by a heuristic perceptual try-and-error process. On the opposite, a library of segmental units may be recorded and segmented in a reasonable lapse of time. However, discontinuities of parameters then usually occur at the junctions between consecutive units, which are unpleasant to the ear. Moreover, these units are rigidly represented as an enumeration of successive control parameters and prove to be very hard to handle successfully for natural-sounding prosodic modifications.

Temporal decomposition can be seen as a technique for structuring a speech segment by setting off different phases in its spectral evolution, as well as describing the way these phases chain together.

Incidentally, a glottal excitation provides a parametric representation of the source characteristics, which is all the more useful in speech synthesis as the parameters are directly related to production phenomena and can therefore be easily controlled.

Speech synthesis by structured segments relies on both these tools : segmental units are first analysed by a glottal-LPC modeling which provides both a parametric excitation and a (glottal) LPC spectrum. This spectrum is then described by a temporal decomposition. Each segmental unit is thus structured into a set of targets, a temporal pattern and a parametric excitation.

Structured segments can be then easily handled by rules; especially when dealing with prosodic factors.

Time distortions can be applied to temporal patterns (i.e. to interpolation functions), in order to render global speaking-rate modifications and local rhythm variations. Typically, stable portions (signed off by a unique \emptyset -function) will be more affected by the rescaling than zones of transition (overlapping \emptyset -functions); this process is more realistic than a global linear time compression or expansion. A temporal realignment has then to be operated on glottal parameters.

Intensity of voiced sounds can be controlled straightforwardly by changing amplitude parameter A of glottal pulses, and thus handle accentuation phenomena (as far as correlative voice texture modifications are neglected).

Pitch is dealt with by reducing or lengthening the closed glottis phase of the pulses, without changing the shape of the open glottis part and while keeping constant the average energy of excitation. A new pitch T thus involves an immediate pro rata adjusting of relative opening duration t_o , of relative closing duration t_c and of amplitude A.

Voice texture mostly depends upon the shape of the glottal wave. This aspect can be taken into account by deforming glottal pulses, that is modifying relative durations t_o and t_c . Adding noise to them or even suppressing some of them may be called for, to create alternate styles of voices like breathy or creaky voice, which often occur locally in natural speech.

Spectral transformations of targets [7] might contribute to articulation quality modifications (careful vs relaxed articulation, rounding, etc...); but this has not been yet thoroughly investigated.

The first step in speech synthesis by structured segments therefore consists in concatenating temporal patterns of segmental units, after having distorted each of them according to duration rules. Each interpolation function is then assigned its original corresponding target. However, at the border of segments, the mean-value of targets associated to the connecting \emptyset -functions can be taken as a unique target for the resulting complete interpolation function. Spectral transformations are possible at that level. A synthetic spectrum is finally reconstructed by simple linear combinations.

Original glottal parameters then undergo a time rescaling in accordance with the duration distortions of the previous step. Those parameters might also be modified in order to fit some former spectral transformations, when rules exist to express the coupling between source and vocal tract.

Segments of parametric excitation are then concatenated, and the resulting glottal parameters are modified in order to impart a stress pattern and a melodic contour to the whole sentence. A glottal temporal waveform is finally regenerated.

Filtering this excitation by the reconstructed spectral parameters provides synthetic speech.

Experiments :

A complete speech synthesizer using structured segments was not constructed yet, as it calls for different topics, the integration of which in a unique system is not straightforward and needs therefore further developments.

Nevertheless, experimental simulations were carried out on a VAX in order to check the validity of the concept.

A first preliminary set of experiments consisted in modifying glottal parameters supplied by a glottal-LPC analysis and next reconstructing the partly synthetic speech. The object of this experiment was to test the resistance of the technique as regards changes of the source characteristics.

Both global shifts and local variations of pitch were experimented and proved to be successful : for instance, an affirmative sentence can be turned into an interrogative one without any noticeable alteration of the speech quality, just by editing and designing again the pitch parameter (by reproducing the melody of an other natural sentence, under the circumstances).

A general consequence of this result is that a glottal excitation is highly desirable for speech synthesis; even for purely segmental systems, since it can be automatically extracted from natural speech (under strict but feasible recording conditions).

An other set of experiments was to construct a few sentences by concatenation of segmental units. These units were analysed beforehand by glottal-LPC and their spectrum was submitted to a temporal decomposition, in order to reproduce the scheme of synthesis by structured segments. Intonation patterns arising from natural sentences were laid on the synthetic ones.

As segmental units, we chose polysyllables (polysounds), i.e. combinations of phonemes including transient and variable sounds such as liquids, semi-vowels or any existing combination of them. These phones prove to be often very different at diphone borders and typically give rise to highly overlapping \emptyset -function patterns. Including them into the set of segments allows to pass round this drawback, at the expense of an increased number of units (approximately a factor of 3).

This experiment supplied good quality intelligible speech, free of the usual buzziness encountered with classical LPC, though it was clearly deteriorated with respect to sentences synthesized from the same units by a pure segmental technique and a glottal excitation. The most obvious defect is a muffling of the transitions, which gives the impression of a recurrent fading of the speech. This decrease in quality is likely to originate mostly from the enlarging of formant bandwidths that L.A.R. interpolations produce, with regard to original transitions. Investigations of more adequate spectral parameters from this point of view is a next step. Nevertheless, this defect may be preferred to classical LPC ones, all the more since it does not increase with prosodic modifications. The glottal-LPC model reduces the dependency between source and vocal tract estimates and is therefore less sensitive to approximations of spectral parameters.

Towards rule-based synthesis :

Rule-based synthesis systems, as the MITalk for example [8], call for different kinds of acoustic knowledge on speech : mainly, a collection of ideal spectral targets representing each phone, allophone or phase of phone, a set of typical interpolation patterns for transitions and some ability to drive jointly a source (i.e. to determine its characteristics and to generate its evolution). Using this knowledge, a succession of synthetic control parameters can be reconstructed from a phonetic string and a set of prosodic marks, and finally be sent to a decoder.

Some experiments were carried out in order to generate (semi-) automatically some parts of this knowledge. A dictionary of allophonic spectral targets was thus created, as well as a typology of temporal patterns.

For that purpose, a library of diphones, uttered by a single speaker, were analysed with the temporal decomposition technique.

Approximately 2400 spectral targets were obtained and manually labelled with a phonemic symbol (in opposition to "phonetic"); except for transition targets for which a diphonemic symbol was used. Targets were grouped afterwards in different subsets, after their phonemic label. Hierarchical cluster trees were built, using the centroid method. From these trees, a number of allophones was decided for each phoneme. The nearest centroid sorting technique was finally used to cluster the observations and provide an average representant for each allophone. Transitions were also clustered into only 7 different classes [9].

Vowels proved to be unimodal, while most consonants led to a voiced (or unvoiced) allophone; some liquids or semi-vowels revealed 3 variants.

Several elements extracted from the dictionary of allophonic spectral targets are shown on figure III.

A very strict protocol showed a good self-consistency, when identifying the original 2400 targets with the 71 elements arising from the dictionary. In spite of the keeping of transitions in the test corpus, and although the following confusions were counted as mistakes :

- mid-opened and mid-closed vowels,
- semi-vowels and their vocalic homologs,
- silent parts of voiceless plosives and
- unvoiced consonants and their voiceless sonant

homologs (or the opposite),

77 % of correct phonemic labels were present in the 3-NN lattice, whereas 53 % were found in first position. Many errors resulted from the confusions quoted above.

The observation of \emptyset -function patterns for diphones made it possible to create a typology of typical temporal structures for transitions between phones. For instance, combinations of neighbouring vowels in the vocalic triangle usually require 2 interpolation functions, while distant ones call for an extra (contextual) event. Diphones consisting of a voiced fricative followed by a vowel are generally modeled with only 2 functions except those with the palatal fricative for which an intermediate function is nearly always found.

Figure IV shows examples of typical temporal structures.

A rough typology of approximately 60 temporal structures was thus listed. Its adequacy was evaluated by counting the number of diphones that agreed with this very typology. When putting aside the 21 % of coarse failures in the temporal decomposition algorithm (splitting of a stable zone, grouping of differing phonemes, missing function, high reconstruction error, ill-formed decomposition), 72 % of temporal structures appeared to be fully predictable; this figure falls to 57 %, when the failures are taken into account (absolute predictability). However, a procedure is possible to impose a temporal pattern to a speech segment and search for the best decomposition under this constraint, which is a generalization of the iterative refinement process [9].

The creation of a dictionary of allophonic spectral targets and the constitution of a typology of temporal structures lay the foundations of a semi-automatical rule design system for speech synthesis. In this framework, glottal wave extraction provides a phonetically relevant parametric description of the excitation. Nevertheless, a prior development is a modeling of source evolution in order to extract rules from it.

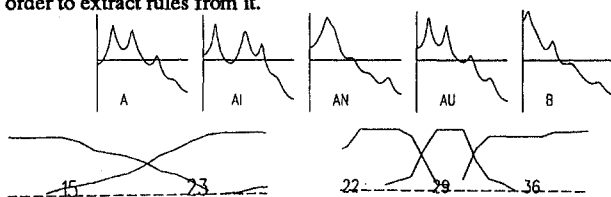


Figure III (top) : some allophonic spectral targets. allophones a, e, ä, ɔ, b, b exp, j, voiced j, d, d exp.

Conclusions :

Speech synthesis by structured segments uses the temporal decomposition technique and a glottal excitation as tools for tending to describe speech in accordance with its underlying phonetic content. The consequences are an increased flexibility of the method and possible issues towards an automatical rule design system.

Some further studies have to be devoted to the search of a more suitable spectral parametrization as regards linear interpolations, in order to improve synthetic speech quality. A model of source evolution is to be proposed and experimented, so that a next step towards rule-based synthesis is overcome.

The temporal decomposition technique itself needs to be enriched; for instance, complemented by a parallel running rupture modeling which would render accidents in the signal [10], while linear interpolations would still be resorted to for describing slowly varying transitions. The phonetic relevance of alternative spectral modeling techniques have also to be investigated; in particular the AR-vector modeling technique [11].

Acknowledgments :

We wish to thank Per Hedelin, Head of the Department "Information Theory" in Chalmers University, Göteborg, Sweden, for having provided us with some pieces of SAP speech analysis software, and we hope that our collaboration will continue through new scientific exchanges.

References :

- [1] G. FANT : *Speech sounds and features*. MIT Press 1973.
- [2] S.M. MARCUS, B.S. ATAL : *Decoding the speech code*. JASA 1986. vol. 80, suppl. 1, S 17.
- [3] B.S. ATAL : *Efficient coding of LPC parameters by temporal decomposition*. ICASSP 1983. pp. 81-84.
- [4] F. BIMBOT, G. CHOLLET, P. DELEGLISE, C. MONTACIE : *Temporal decomposition and acoustic-phonetic decoding of speech*. ICASSP 1988. vol. 5, pp. 445-448.
- [5] G. FANT : *Voice source dynamics*. STL-QPSR. KTH 1980. vol. 2-3.
- [6] P. HEDELIN : *High quality glottal-LPC vocoding*. ICASSP 1986. pp. 465-468.
- [7] C. MONTACIE, K. CHOUKRI, G. CHOLLET : *Speech recognition using temporal decomposition and multi-layer feed-forward automata*. ICASSP 1989. vol. S1, pp. 409-412.
- [8] J. ALLEN, M.S. HUNNICUT, D. KLATT : *From text to speech : the MITalk system*. Cambridge University Press 1987.
- [9] F. BIMBOT : *Synthèse de la parole : des segments aux règles, avec utilisation de la décomposition temporelle*. Thèse de Doctorat E.N.S.T. 1988.
- [10] M. BASSEVILLE, A. BENVENISTE : *Detection of abrupt changes in signals and dynamical systems*. Springer-Verlag 1986.
- [11] A. DE LIMA VEIGA, Y. GRENIER : *A multi-step excited model for speech parameter trajectories*. Submitted to IEEE Trans. ASSP 1988.

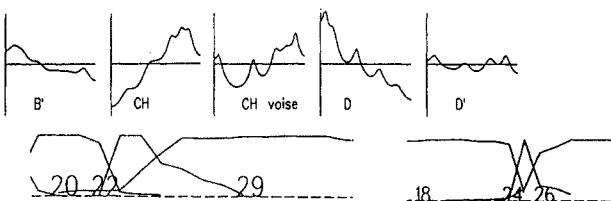


Fig IV (bottom) : A few typical temporal structures. neighbouring vow, distant vow, son plos + vowel, unv fric + vowel.