Speech Analysis with Production Constraints

Abstract of Ph.D. Thesis, by **Sacha Krstulović**

December 2001

State of the art speech analysis and feature extraction techniques rely mainly on simplified auditory models (e.g. the Mel scale or PLP features). These systems can model any sound and are not particularly specialized in the modeling of speech. Hence, they fail to reflect some typical speech characteristics such as co-articulation.

To bridge this gap, we propose to match some speech production paradigms with Automatic Speech Processing technologies (ASP). This matching is based *(1)* on an analogy between the Linear Prediction (LP) of speech and the acoustic modeling of lossless tubes, and *(2)* on the fact that most of today's state of the art speech production models use an acoustic tube as the interface between the acoustic level and the speech production strategy level. In this framework, we develop two innovative feature extraction methods: the Non-Uniform Topology (NUT) analysis, and the "Relating Acoustics to a Linear Shape Model" (ReALiSM) method.

To establish the **NUT analysis**, we begin with generalizing the traditional LP lattice filter/lossless tube equivalence to the case of tubes discretized in unequal-length sections, such as the non-uniform tube used for the Distinctive Regions and Modes (DRM) speech production model.

We show that imposing unequal-lengths to the tube sections is equivalent to constraining some reflection coefficients to stay zero-valued in the corresponding lattice filter. This "Non Uniform Topology" constraint allows to de-couple the number of degrees of freedom (DoFs) of the model from the dimensions of its acoustic counterpart (given by the number of poles). To use this new model as an analysis tool, we derive some relevant parametric estimators, based on the analytic minimization of a well-defined error criterion. Finally, remarking that a fixed non-uniform topology (e.g., the one of the DRM production model) may not be optimal for every part of speech, we propose a method to optimize the repartition of the lengths/delays.

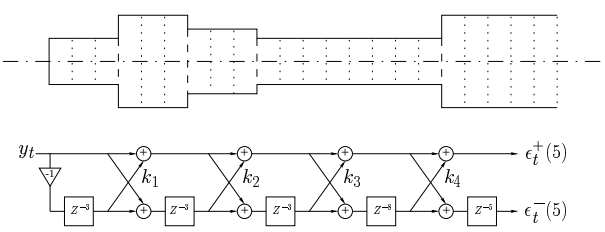
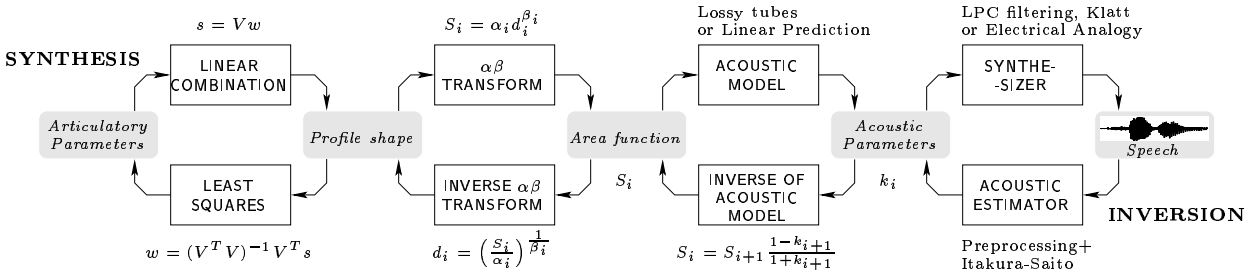


Fig.1: a NUT tube and the equivalent lattice filter.

The assessment phase shows that NUT models permit a significant reduction in the number of parameters necessary to describe a speech spectrum, while keeping a high level of spectral accuracy. It is verified that they consistently produce a lower residual error than the unconstrained filters with and equal number of DoFs. It is also verified that the NUT models are consistent with spectral analysis since the topology optimization helps minimizing the spectral distortion induced by the reduction of the number of DoFs. Moreover, the tube topologies themselves may be used as an analysis tool.

Alternately, to establish the **ReALiSM method**, we implement a projection of the solution of inverse LP lattice filtering into the parameter space of Maeda's linear vocal tract shape model, through a series of linear and non-linear transformations:



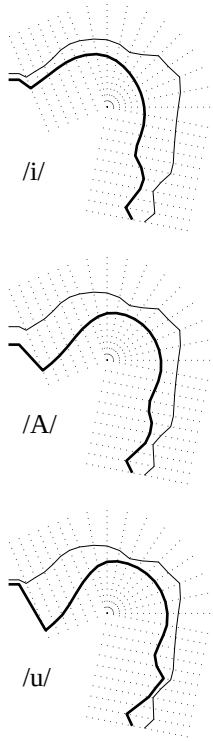


Fig.2: Inversion of human vowels.

The assessment of this system is realized in two phases. First, it is verified that information losses that occur within Least-Squares smoothing, area functions resampling and reflection coefficients estimation still allow for recovery of synthetic template tract shapes. A set of synthetic vowels is produced (cardinal French vowels registered in the UPSID phonemic database), and informal listening tests ensure that they are acceptable despite the fixed length approximation and the lossless LPC synthesizer. Subsequent inversion results show that the estimated shapes are close to the original synthetic shapes.

In a second phase, the system is used to invert real speech recorded from a French male speaker in a quiet environment. Several vowel sequences and VCV sequences are tested. For example, results corresponding to the vowels /i A u/ are given on figure 2. The system locates cavities at phonetically relevant places of articulation (e.g. front for /A/, back for /i/). Lip apertures are also realistic, e.g., in an /A b i/ sequence the /b/ consonantal closure is detected. This system offers significant advantages over existing acoustico-articulatory inversion schemes, such as real-time computation, modularity and links with Digital Signal Processing techniques.

Since Linear Prediction analysis is used as a feature extraction method in most of the main ASP technologies (such as speech coding, speech/speaker recognition, speech synthesis, speech enhancement etc.), the production-based analysis methods that we have developed create a new gateway for the **integration of speech production constraints in the main classical ASP applications**. To assess the benefits that these methods introduce, we propose multiple ways to exploit the NUT and ReALiSM systems in various branches of the ASP domain. In particular, we provide and discuss encouraging preliminary speech recognition results.