ARTICULATORY SYNTHESIS USING A STOCHASTIC TARGET MODEL OF SPEECH PRODUCTION

Gordon Ramsay and Li Deng
Dept. of Electrical & Computer Engineering, University of Waterloo, Canada.

ABSTRACT

A stochastic target model for articulatory synthesis is described, where articulator motion is modelled by a linear system driven by random target functions and modulated by a finite-state Markov model. States in the model represent overlapping phonological units, while probability distributions for the associated target regions represent systematic articulatory variation. Simple examples of random synthetic speech are given.

INTRODUCTION

Speech recognition and synthesis have traditionally been approached through entirely different methodologies; the former based largely on *trainable models* which reflect the statistical characteristics, but not the mechanisms, of real speech; the latter based on *rule-driven systems* which incorporate a great deal of a-priori knowledge about speech mechanisms, but which lack the ability to adapt automatically to a particular corpus of data. In neither of these approaches is it usually possible to model articulatory phenomena directly, due to the lack of an appropriate articulatory representation.

In a previous presentation, a framework for articulatory speech recognition was outlined, based on a stochastic target model of speech production constructed around explicit articulatory and acoustic models of the vocal tract [1]. It was shown that the parameters of the model can, in theory, be trained automatically from a corpus of acoustic data using the EM algorithm.

In this paper, it is shown that the model can also be used for speech synthesis by sampling the underlying probability space. The resulting output incorporates a degree of non-deterministic but systematic variation, reflecting some of the possibilities for compensatory phenomena observed in real speech.

Simple examples of $(VCV)^+$ utterances are generated to demonstrate that the model is capable of producing plausible articulator and formant trajectories automatically.

MODEL STRUCTURE

Assume an underlying probability space $(\Omega, \mathcal{F}, P)$, and let $S = \{S_m : m \in \mathcal{N}\}$ be a finite-state Markov chain taking values in $\mathcal{S} = \{s_i : i = 1 \ldots N\}$, with transition matrix $\Pi = [\pi(i,j) : i,j \in \mathcal{S}]$, where $\pi(i,j) = P(S_{m+1} = j | S_m = i)$ and $\sum_j \pi(i,j) = 1$. Each state in the Markov chain represents a phonological symbol or, more generally, a combination of overlapping symbols, while any path through the state structure generates the symbol sequence for a particular utterance.

To describe the temporal characteristics of each phonological sequence, define a second process $T = \{T_m : m \in \mathcal{N}\}$. Each T_m represents the number of time frames spent in state S_m, and is assumed for convenience to be Poisson-distributed with parameter $\mu_\tau(S_m)$ drawn from a set $\mathcal{T} = \{\mu_\tau(i) \in \mathcal{R} : i \in \mathcal{S}\}$ according to the Markov state. The T_m are conditionally independent given S.

Each phonological symbol is assumed to possess a number of underlying physical correlates, which may be articulatory, acoustic or perceptual in nature. The fundamental modelling assumption is that the set of correlates for each symbol can be projected onto an equivalent *target region* in a Euclidean space of articulatory parameters $\mathcal{X} = \mathcal{R}^p$. The target region associated with any individual symbol can then be modelled as a distribution function on $\mathcal{X}$, giving the probability that any particular vocal tract configuration in $\mathcal{X}$ is capable of realizing the phonetic correlates associated with that symbol. Every time a transition occurs in the Markov chain, a new target configuration is chosen according to the target distribution for the new state, and held constant until the next state transition occurs.

To represent this, define a target process $U = \{U_m : m \in \mathcal{N}\}$ taking values in $\mathcal{X}$, where the U_m are independent conditioned on S, and each U_m is Gaussian-distributed with mean $\mu_u(S_m)$ and covariance matrix $\Sigma_u(S_m)$, selected from a set of target parameters $\Theta = \{(\mu_u(i) \in \mathcal{R}^p, \Sigma_u(i) \in \mathcal{R}^{p \times p}) : i \in \mathcal{S}\}$ according to the Markov state S_m. The extension to arbitrary continuous distributions on $\mathcal{X}$ is straightforward by approximation using Gaussian mixtures, and more complicated parameterizations are clearly possible where the target distributions interact or are made to vary with time.

The processes S,T,U then describe the generation of a random distribution of vector-valued target functions in articulatory space for a class of phonological state sequences.

Now, let $X = \{X_n : n \in \mathcal{N}\}$ be a random process on $\mathcal{X}$ representing the articulatory state, and let $Y = \{Y_n : n \in \mathcal{N}\}$ be a measurement process generating observations of X in an acoustic space $\mathcal{Y} = \mathcal{R}^q$. Assume that the initial state X_1 is distributed as $N(\mu_1 \epsilon \mathcal{R}^p, \Sigma_1 \epsilon \mathcal{R}^{p \times p})$ and define zero-mean Gaussian i.i.d. processes $V = \{V_n : n \in \mathcal{N}\}$ and $W = \{W_n : n \in \mathcal{N}\}$ to represent unmodelled perturbations in $\mathcal{X}$ and $\mathcal{Y}$, with covariance matrices $\Sigma_{vv} \epsilon \mathcal{R}^{p \times p}$ and $\Sigma_{ww} \epsilon \mathcal{R}^{q \times q}$ respectively.

Assume furthermore that X evolves in time according to the linear difference equation (1) driven by U (cf. [2][3]), and that Y is generated from X through a memoryless non-linear transformation $h : \mathcal{X} \rightarrow \mathcal{Y}$ as seen in (2).

Here d is the order of the system, and the matrices $A_i(j) \in \mathcal{R}^{p \times p}$ are selected from a set of system parameters $\mathcal{A} = \{A_i(j) : i = 1 \ldots d, j \in \mathcal{S}\}$ according to S_m. Since each control state S_m influences T_m frames of the articulatory process X, a random index function $J : \Omega \times \mathcal{N} \rightarrow \mathcal{N}$ is needed to cross-reference points in (S,T,U) and X.

$$X_{n+1} = \sum_{j=1}^{d-1} A_j(S_{J(n)}) X_{n+1-j} + A_d(S_{J(n)}) U_{J(n)} + V_n, \quad (1)$$

$$Y_n = h(X_n) + W_n. \quad (2)$$

This completes the description of the overall model structure. The function $h(.)$ represents the articulatory-acoustic mapping, and can be approximated using a codebook of points simulated from an acoustic model of the vocal tract. Provided that the phonetic correlates chosen for each phonological state can be expressed in terms of quantities which can be measured from model simulations, the corresponding target distributions can easily be derived from the codebook by defining an appropriate normalized cost function on articulatory space. Initial estimates for the duration parameters and the time constants of the state recursions can be measured from acoustic or articulatory data.

SIMULATION RESULTS

The model can now be used for speech synthesis by randomly generating sample paths from the probability space, using a Monte-Carlo technique. A state path through the Markov chain is first selected according to the transition matrix. Once the state path has been chosen, corresponding durations and target points are generated as a sequence of independent random variables with distributions determined by S. The articulatory state is Gaussian conditioned on (S,T,U), with mean and covariance that can be calculated recursively from the initial distribution for X_1, the target sequence U, and the sequence of system matrices defined by the Markov state path. Once the distribution of X is known, synthetic speech

can be obtained by generating a single random sample path of X and passing the result through the acoustic model.

Figure 1 shows a simple state structure representing $(VCV)^+$ utterances for the vowels /a/,/i/,/u/ and consonants /r/,/d/,/w/. Associated with each state is a target region representing the appropriate oral stricture. For the vowels, the region is characterized by a set of constraints on the first two formants, for example $/a/ = \{600 < F1 < 1000, 1000 < F2 < F1 + 500\}$, $/i/ = \{F1 < 400, F2 > 2000\}$, $/u/ = \{F1 < 400, F2 < 1000\}$ (in Hz), together with a requirement that the formant energy be greater than a sonorant threshold. For consonants, a mixture of acoustic and articulatory correlates are used. /r/ is defined by $\{F3 - F2 < 400, F2 < 1900\}$, /d/ by a closure along the alveolar ridge with $\{1600 < F2 < 1900\}$, while /w/ requires protrusion of the lips with $\{F1 < 300, F2 < 500\}$. The states are intended to represent combinations of abstract units, but these need not necessarily be segmental (cf. [4]).

Six parameters of a version of Mermelstein's model [5], shown in Figure 2, were chosen to form the dimensions of the articulatory space. A small codebook of 25000 entries generated from a finite-difference solution of the wave equations was used to provide measurements of the formants, average acoustic energy, and constriction location for a uniform distribution of points on $\mathcal{X}$.

The articulatory image of each target was then constructed by fitting a single Gaussian distribution to the class of all points satisfying the appropriate definition. Figure 3 shows two different projections of the sample distribution for /d/, illustrating some of the correlation patterns which arise automatically from this technique.

Figures 4 and 5 show spectrograms of two typical utterances produced from sample paths of X using an articulatory synthesizer, together with the parameter traces for X (−) and U (···). Realistic formant trajectories have been produced using only the statistical properties of target regions derived from relatively abstract and flexible phonetic specifications.

CONCLUSIONS

A stochastic target model for articulatory synthesis has been outlined, based on Monte-Carlo simulation of a Markov-modulated linear system. The model permits a compact articulatory representation of speech in terms of a relatively small number of statistical parameters which, in theory, it should eventually be possible to train from a corpus of acoustic data. In conjunction with existing filtering algorithms, the same modelling framework may also be used for speech recognition. Simple examples of synthetic VCV utterances have been constructed, and demonstrate that the model is indeed capable of reproducing many of the characteristics of real speech, although the quality does not at present approach that of formant synthesis. Future work will concentrate on improving the underlying model and adapting it to real data.

REFERENCES

[1] Ramsay G., Deng L., (1994) "A Stochastic Framework for Articulatory Speech Recognition," *JASA* 95 (5) Pt.2 Abstract 2aSP19.

[2] Saltzman, E. L., Munhall, K. G. (1989), " A dynamical approach to gestural patterning in speech production," *Ecological Psychology* 1 (4) pp. 333-382.

[3] Shirai K., Honda M., (1976) "Estimation of articulatory motion," in *Dynamic Aspects of Speech Production*, University of Tokyo Press.

[4] Browman C.P., Goldstein L., (1992) "Articulatory phonology : an overview," *Phonetica* 49 pp. 155–180.

[5] Rubin, P., Baer, T., Mermelstein, P. (1981), "An articulatory synthesizer for perceptual research," *JASA* 70 pp. 321-328.

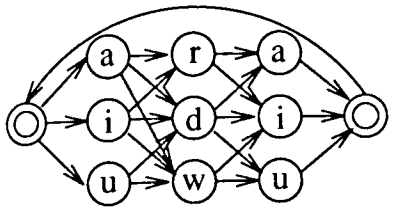

Figure 1: $(VCV)^+$ state structure.

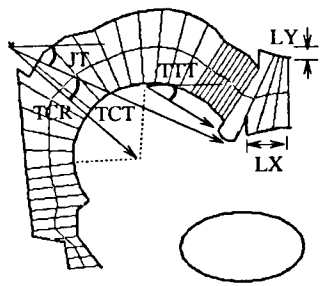

Figure 2: Articulatory model.

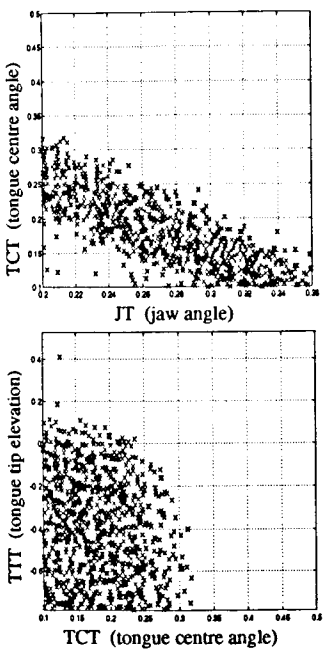

Figure 3: /d/-target projections.

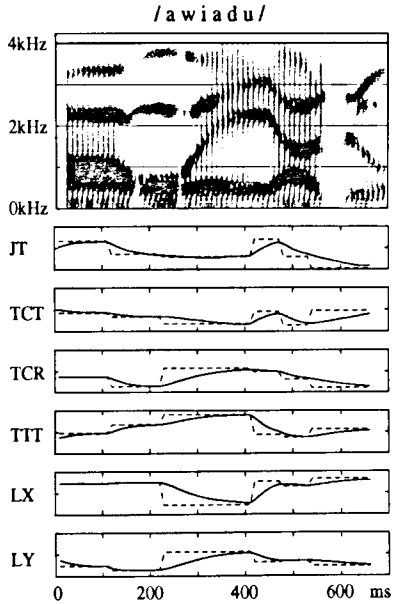

Figure 4: Random synthetic speech

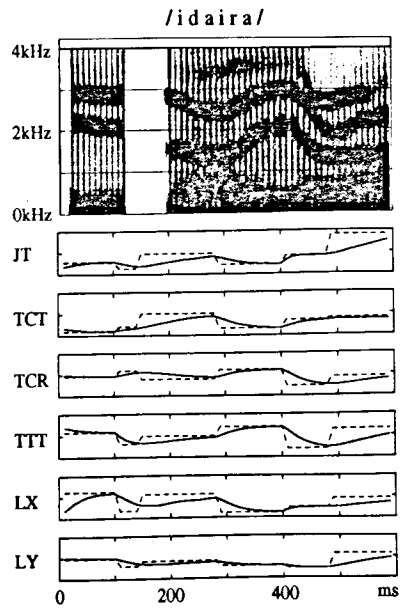

Figure 5: Random synthetic speech