NONLINEAR DYNAMICAL SYSTEM BASED ACOUSTIC MODELING FOR ASR

Narada D. Warakagoda and Magne H. Johnsen

Department of Telecommunications
NTNU, O.S. Bragstad Plass 2B
N-7034, Trondheim, Norway
warakago,mhj@tele.ntnu.no.

ABSTRACT

The work presented here is centered around a speech production model called Chained Dynamical System Model (CDSM) which is motivated by the fundamental limitations of the mainstream ASR approaches. The CDSM is essentially a smoothly time varying continuous state nonlinear dynamical system, consisting of two sub dynamical systems coupled as a chain so that one system controls the parameters of the next system. The speech recognition problem is posed as inverting the CDSM, for which we propose a solution based on the theory of Embedding. The resulting architecture, which we call Inverted CDSM (ICDSM) is evaluated in a set of experiments involving a speaker independent, continuous speech recognition task on the TIMIT database. Results of these experiments which can be compared with the corresponding results in the literature, confirm the feasibility and advantages of the approach.

1. INTRODUCTION

From the statistical pattern recognition point of view, ASR is a classification problem where the vector to be classified is the input speech waveform $\mathbf{s}$ and classes correspond to different sentences $\mathbf{W}$. The ASR system itself should represent the joint probability $p(\mathbf{s}, \mathbf{W})$, but the common practice is to focus the attention on $p(\mathbf{O}, \mathbf{W})$, where $\mathbf{O}$ is the feature vector sequence, rather than $p(\mathbf{s}, \mathbf{W})$itself. The missing link, $p(\mathbf{s}|\mathbf{O})$, is usually modeled as a deterministic relationship between $\mathbf{s}$ and $\mathbf{O}$ and commonly known as feature extraction. This approach as practiced today has at least two main drawbacks. Firstly, the connection between $\mathbf{s}$ and $\mathbf{O}$ are made through procedures such as Fourier analysis, which are based on the linear philosophy [1]. It is highly questionable how such a simplified view of a nonlinear phenomenon can serve the purpose of modeling $p(\mathbf{s}|\mathbf{O})$. Secondly the “artificial” nature of $\mathbf{O}$ due to this approach will not allow easy and efficient modeling of $p(\mathbf{O}, \mathbf{W})$ using, for example, the predictive relationship between successive feature vectors [2]. The lack of such a possibility has contributed the errorness assumption of feature vector independence to become a de-facto standard in acoustic modeling.

One obvious solution to the above problems is to look for a modeling paradigm which directly operates on the waveform space and respects the inherent nonlinear nature of the process. Nonlinear dynamical systems fit nicely to this requirement and in fact, several authors have tried to apply these techniques on speech signals (see [3] and references therein). However these techniques work properly only for long stationary segments of signals.

In this paper, we propose a nonlinear dynamical systems based acoustic modeling approach, in which the dynamics within stationary segments as well as transition dynamics are taken care of. The core of the approach is to view the human speech production system as two dynamical systems coupled as a chain. In this paper we refer to this as the Chained Dynamical System Model (CDSM). The input to the CDSM is an abstract code representing the speech unit sequence (eg: a phoneme sequence) which is corresponding to the speech waveform at the output. Therefore the ASR problem can be viewed as inverting the CDSM. We propose a solution to this inversion problem, which is rooted in Taken’s embedding theorem [4].

The rest of the paper is organized as follows. In section 2 we outline the structure of the CDSM. In section 3, how the CDSM can be inverted is sketched and section 4 is devoted to a brief description of the training procedure for the ICDSM. Next, experiments and results are presented in section 5. Finally, in section 6, we make some concluding remarks on the work.

2. THE CDSM

Figure 1 depicts the system what we call the CDSM. Here, the nonlinear function $\boldsymbol{\mathcal{F}}_2(\cdot)$ models the dynamics of the articulator configurations contained in the *state vector* $\boldsymbol{\varkappa}_2(k)$. These dynamics are under supervision of the *control vector* $\boldsymbol{a}(k)$ which represents the current sound class (eg: phoneme) to be generated. We denote this dynamical system by $\mathbf{\Phi}_2$. The nonlinear functions $\boldsymbol{\mathcal{F}}_1(\cdot)$ and $h(\cdot)$ model the bio-mechanics of the production of the speech signal $s(k)$ having $\boldsymbol{\varkappa}_1(k)$ as the state vector. We call this dynamical system $\mathbf{\Phi}_1$.

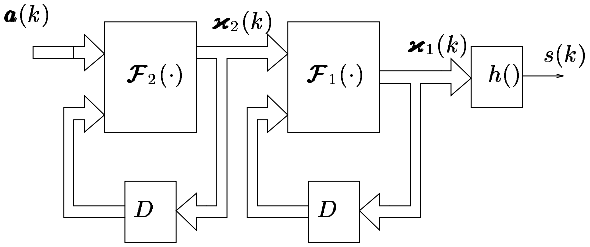
$\boldsymbol{a}(k)$
$\boldsymbol{\mathcal{F}}_2(\cdot)$
$\boldsymbol{\varkappa}_2(k)$
$\boldsymbol{\mathcal{F}}_1(\cdot)$
$\boldsymbol{\varkappa}_1(k)$
$h()$
$s(k)$
D
D

Fig. 1. *Chained Dynamical System model (CDSM). Here D denotes a delay element.*

3. INVERSION OF THE CDSM

With reference to the CDSM of speech production, the speech recognition problem can be posed as finding the control vector sequence $\boldsymbol{a}(k),\ \ k = 0, 1, \ldots$ for a given speech waveform $s(k),\ k = 0, 1, \ldots$. This is nothing else but inverting the CDSM. Our solution to this inversion problem is based on a generalized version of Taken's theorem [4], according to which a frame of waveform samples

$$\mathbf{x}_1(k) = [s(k), s(k-\tau), s(k-2\tau), \ldots, s(k-(m-1)\tau)]$$

has a one-to-one smooth correspondence with the actual state $\boldsymbol{\varkappa}_1(k)$ of the dynamical system $\mathbf{\Phi}_1$, if m is greater than twice the dimension d of the actual state space of $\mathbf{\Phi}_1$. However, there is no theoretical way to obtain d or τ. Fortunately, there are many practical recipes available to cover this deficiency [5]. But these recipes are ad-hoc in nature and not optimal with respect to a relevant criterion. Therefore as suggested in [6], we view $\mathbf{x}_1(k)$ as a projection of a sufficiently long waveform vector on the m-dimensional space. That is

$$\mathbf{x}_1(k) = \mathbf{P}(\hat{\mathbf{v}}(k)) \tag{1}$$

where $\mathbf{P}(\cdot)$ is the (possibly nonlinear) projection operator and $\hat{\mathbf{v}}(k) = [s(k), s(k-1), \ldots, s(k-\hat{K}_f+1)]^t$ is a waveform vector of length $\hat{K}_f$ which is sufficiently high. Note however that $\mathbf{P}(\cdot)$ has to estimated using available training data, which is most effective when we have rough values for m and τ (and hence $\hat{K}_f$) at hand from ad-hoc approaches.

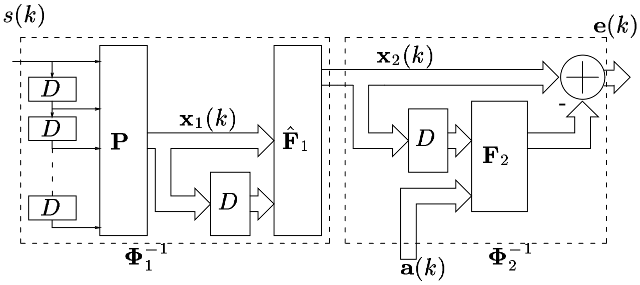

Fig. 2. *Inversion of the CDSM.*

Once we have an estimate $\mathbf{x}_1(k)$ for $\boldsymbol{\varkappa}_1(k)$, we can find an estimate for $\boldsymbol{\varkappa}_2(k)$, using the relation $\boldsymbol{\varkappa}_1(k) = \boldsymbol{\mathcal{F}}_1(\boldsymbol{\varkappa}_1(k-1), \boldsymbol{\varkappa}_2(k))$ as seen from figure 1. Namely, assuming that $\boldsymbol{\mathcal{F}}_1(\cdot)$ is invertible we can write

$$\mathbf{x}_2(k) = \hat{\mathbf{F}}_1(\mathbf{x}_1(k), \mathbf{x}_1(k-1)) \tag{2}$$

where $\mathbf{x}_2(k)$ and $\hat{\mathbf{F}}_1(\cdot)$ are estimates for $\boldsymbol{\varkappa}_2(k)$ and $\boldsymbol{\mathcal{F}}^{-1}(\cdot)$ respectively.

Finally, an estimate $\mathbf{a}(k)$ for the control vector $\boldsymbol{a}(k)$ can be obtained through the relation $\boldsymbol{\varkappa}_2(k) = \boldsymbol{\mathcal{F}}_2(\boldsymbol{\varkappa}_2(k-1), \boldsymbol{a}(k))$ which is observable from figure 1. As was done for the case of $\boldsymbol{\mathcal{F}}_1(\cdot)$, we can assume that $\boldsymbol{\mathcal{F}}_2(\cdot)$ is invertible and obtain $\mathbf{a}(k)$ right away. But since we would like to avoid such assumptions as much as possible, a more demanding, but more accurate procedure is followed. Namely, first it is assumed that an estimate $\mathbf{F}_2(\cdot)$ for $\boldsymbol{\mathcal{F}}_2(\cdot)$ and $\mathbf{a}(k)$ are errornously known. That is

$$\mathbf{x}_2(k) = \mathbf{F}_2(\mathbf{x}_2(k-1), \mathbf{a}(k)) + \mathbf{e}(k) \tag{3}$$

Then parameters of $\mathbf{F}_2(\cdot)$ as well as $\mathbf{a}(k)$ are adjusted in such a way that the norm of $\mathbf{e}(k)$ is minimized.

The whole procedure of inversion as defined by eqns. 1, 2 and 3 is presented pictorially in figure 2. The architecture shown in figure 2 can be further refined with regard to two aspects.

First, the architecture shown in figure 2 uses the same *time constant* T_s in both dynamical systems $\mathbf{\Phi}_1$ and $\mathbf{\Phi}_2$. (i.e. in both systems, predictive relationships of the state vectors which are T_s time units apart are utilized). But it is generally accepted that articulator configuration dynamics (eg: vocal tract dynamics) takes place in a slower time scale than does the within-articulator dynamics. Therefore, time constant T_{s2} of $\mathbf{\Phi}_2$ must be larger than that ($T_{s1} = T_s$) of $\mathbf{\Phi}_1$. This change can be achieved by inserting a *decimator* between $\mathbf{\Phi}_1^{-1}$ and $\mathbf{\Phi}_2^{-1}$ in the architecture in figure 2. We select a decimator which averages the past $\left[\frac{T_{s2}}{T_{s1}}\right]$ samples in every T_{s2} time for this purpose. With such a decimator, we can show by simple block manipulation that the mapping from $\hat{\mathbf{v}}(k)$ to $\mathbf{x}_2(k)$ (performed by the projection block $\mathbf{P}$ and the block $\hat{\mathbf{F}}_1$) is equivalent to a mapping from a longer speech frame $\mathbf{v}(k)$ to $\mathbf{x}_2(k)$ by a single function $\mathbf{F}_p$. Here $\mathbf{v}(k)$ is given by

$$\mathbf{v}(k) = [s(kk_D), s(kk_D-1), \ldots, s(kk_D-K_f+1)]$$

where $k_D = \frac{T_{s2}}{T_{s1}}$ and $K_f = k_D + \hat{K}_f$.

The second refinement is to consider the error quantity $\mathbf{e}(k)$ a random variable with a zero mean Gaussian distribution. We accomplish this by introducing a block $\mathbf{Pr}_j$ for each class j whose output is given by

$$p(k,j) = (2\pi)^{-\frac{D_E}{2}} |\mathbf{\Sigma}_j|^{-\frac{1}{2}} \exp\left\{ -\frac{1}{2}\mathbf{e}(k)^t \mathbf{\Sigma}_j^{-1} \mathbf{e}(k) \right\} \tag{4}$$

where D_E is the dimension of $\mathbf{e}(k)$ and $\mathbf{\Sigma}_j$ is the covariance matrix.

With these modifications, our inverted CDSM will appear as depicted in figure 3.

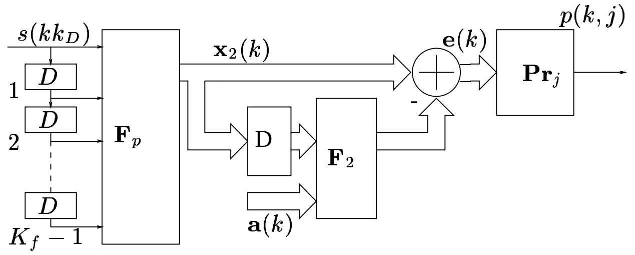

Fig. 3. *A modified version of the ICDSM suitable for practical ASR*

4. TRAINING ALGORITHM FOR THE ICDSM

The goal of training is to estimate the parameters of the function blocks $\mathbf{F}_p$ and $\mathbf{F}_2$ which are implemented as Multilayer Perceptrons, as well as covariance matrix $\mathbf{\Sigma}_j$ and control input $\mathbf{a}_j$ for each class $j = 0, 1, \ldots, C-1$.

We use two types of training algorithms; an isolated mode algorithm which depends on boundary information and a discriminative continuous mode algorithm. In both cases a gradient technique based on the RPROP update rule is employed [7].

In the isolated mode of training, the objective function used for gradient based optimization is

$$E = -\sum_{\mathcal{T}} \sum_{k=0}^{N-1} \ln\left(p(k, Q(k))\right) \tag{5}$$

where $\mathcal{T}$ is the training set and N is the number of frames of the current utterance. $Q(k)$ is a function which maps the frame index k to the class index j. Based on this equation, it is trivial to obtain the derivative of E with respect to the error probability $p(k, Q(k))$.

The discriminative training algorithm is based on the MMI criterion [7], which is defined considering the whole error probability matrix $p(k, j),\ k = 0, 1, \dots, N_i - 1,\ j = 0, 1, \dots, C - 1$ for each utterance in $\mathcal{T}$. Since we can view this matrix as a trellis (with its elements as nodes), there is a straightforward similarity to the HMM based systems. Therefore the derivative of the criterion with respect to $p(k, j)$ can be obtained using the formula given in [7].

Once the gradient of the objective function with respect to the error probabilities are known, they can be back-propagated through the whole system similar to the approach in [7]. This gives the gradients with respect to system parameters so that those can be updated towards an optimum point.

5. EXPERIMENTS AND RESULTS

In order to evaluate the ICDSM, we select the so called 39-class task on the TIMIT database [7]. The training set for the task is prepared by picking all the SI and SX sentences (3696 sentences all together) from all 462 speakers in the TIMIT training set, while taking the so called *core test set* for testing.

The baseline system we used in these experiments has the raw form as depicted in figure 3. The system is fed with speech frames of 25ms taken at 10ms intervals, which implies that $K_f = 400$ and $k_D = 160$ as the sampling rate is 16kHz. All vectors in the system ($\mathbf{x}_2$, $\mathbf{a}$ and $\mathbf{e}$) are dimensioned to 8. Following the state concept in HMMs, we use three control vectors to represent each class, even though in earlier sections we assume that a single control vector represents a class, for the sake of clarity.

The implementation details of the function blocks are as follows. The function block $\mathbf{Pr}_{j,s}$ for each class j and "state" s is implemented with a diagonal covariance matrix $\mathbf{\Sigma}_{j,s}$. $\mathbf{F}_2(\cdot)$ is implemented as a three layer MLP with dimensions (8-10-10-8), tan-sigmoids in the first two layers and a linear output layer. The inputs to $\mathbf{F}_2(\cdot)$, namely $\mathbf{x}_2(k-1)$ and $\mathbf{a}(k)$ are combined using elementwise multiplications. We studied several different implementations of the function block $\mathbf{F}_p$, because it is a crucial component of the ICDSM.

(1) An MLP like architecture obtained through generalization of a Mel Frequency based Cepstral Coefficient (MFCC) calculation procedure [8]. This architecture called MFCC-MLP is dimensioned to represent the calculation of 8-dimensional MFCCs using a 256-point Fourier Transform based 24-channel filter bank.

(2) The first layer of the MLP-like structure in item 1, is modified to represent the Fast Fourier Transform (FFT) algorithms. In this way a computational advantage is obtained.

(3) A three layer MLP with dimensions (400-24-24-8), tan-sigmoids in the first two layers and a linear output layer. This MLP is initialized randomly using a normal distribution with zero mean and 0.1 variance.

(4) The same MLP in item 3, but initialized in such a way that MFCCs are produced at the output when fed with speech frames.

(5) A globally recurrent network with dimensions (208-24-24-8), tan sigmoids in the first two layers and a linear output layer. The architecture is initialized randomly.

Before training, any element which does not have a systematic method for initialization, is initialized randomly. Then the isolated mode training procedure is run. Finally, discriminative training is carried out in continuous mode. The results for these experiments are shown in table 1, where percentage correct (%Corr) and percentage accuracy (%Accu) are calculated as in [7].

$\mathbf{F}_p(\cdot)$ variant/code	%Corr	%Accu
MFCC-MLP (Freezed $\mathbf{F}_p$)/ VA1	59.38	55.97
MFCC-MLP (Freezed layer 1)/ VA2	62.30	58.72
MFCC-MLP (FFT based layer1)/ VA3	64.75	60.44
MLP (randomly initialized)/ VA4	65.39	61.62
MLP (initialized to MFCC)/ VA5	65.56	62.78
Recurrent Net/ VA6	67.37	64.11

Table 1. Recognition results for the 39-class recognizer, for different architectures representing $\mathbf{F}_p(\cdot)$. VAx is a code name for the $\mathbf{F}_p(\cdot)$ variant

As seen from table 1, the best results is obtained for the recurrent net implementation of $\mathbf{F}_p$. Further the randomly initialized MLP performs extremely well. This means that $\mathbf{F}_p$ implementations which are completely independent of the traditional feature extraction algorithms can perform almost as good as (or even better than) those related to traditional algorithms. Further, we can see that any form of optimization of $\mathbf{F}_p$ gives rise to considerable improvements over the case of unoptimized $\mathbf{F}_p$ (MFCC-MLP with freezed $\mathbf{F}_p$ in table 1).

One drawback of the ICDSM is that it specializes merely on the dynamics in the state space, while ignoring the static information about the state space itself. On the other hand , HMMs incorporate a lot of static information (of MFCCs for example), albeit not of exactly the same state space. Therefore superior results can be expected from a combined system. To study this possibility we use a separately trained single mixture monophone HMM system, which is based on 26-dimensional feature vectors consisting of MFCCs, log energy and their deltas. This HMM is trained using the Conditional Maximum Likelihood criterion as in [7] and in stand-alone testing it gives 64.45 %correct and 61.22 %accuracy. The ICDSM and the HMM system are combined in the testing phase by summing state conditioned likelihoods from the two systems. Results for this experiment are shown in table 2. These results show that we really can gain something by combining these systems. Additional static information brought into the system by the HMMs, in all cases has caused an increase of 3-4% in recognition accuracies.

The systems described so far do not have an explicit mechanism for absorbing speaker variations. One way to incorporate this ability is to use a mixture of predictors (instead of a single predictor) in the ICDSM. In practice this can be achieved by having several control vectors per class, but still using the same single

Code of the $\mathbf{F}_p(\cdot)$ variant	ICDSM		Combined	
	%Corr	%Accu	%Corr	%Accu
VA3	64.75	60.44	66.86	63.69
VA4	65.39	61.62	69.17	65.17
VA5	65.56	62.78	69.55	65.63
VA6	67.37	64.11	70.95	67.86

Table 2. *Recognition results for the 39-class task using the combined ICDSM-HMM recognizer.*

predictor function block $\mathbf{F}_2$. The class conditioned probability can then be expressed as a weighted sum of the mixture component prediction error probabilities.

In order to evaluate the ICDSM with mixture of predictors, the system with $\mathbf{F}_p$ variant VA4 (the one with randomly initialized MLP) is used. Table 3 summarizes the results. It is clear from these results that as the number of mixture components in the predictor increases the results get improved. This behavior can be expected to continue until the advantage of the additional parameters of the added mixture components levels out with the reduction of the generalization ability.

#mixture-Predictors	ICDSM		Combined HMM-ICDSM	
	%Corr	%Accu	%Corr	%Accu
1	65.39	61.62	69.17	65.17
2	67.67	63.31	70.39	66.29
3	69.25	64.61	70.71	67.73

Table 3. *Recognition results for the mixture predictor based ICDSM.*

6. CONCLUDING REMARKS

The experiments and results presented in this paper show that the ICDSM can be successfully used as a speech recognizer. The architectures tested here are characterized by the property of effective extraction of dynamic information, especially those at transitions. Based on this, we can expect that they are capable of filtering out co-articulation effects in an efficient manner. Therefore it would be interesting to compare the performance of the ICDSM with that of systems which are specifically designed for handling co-articulation effects. Table 4 shows the results for some of such systems which are evaluated using the TIMIT-database. Even though these results cannot be compared directly, we see that performance of the ICDSM based systems lies reasonably closely to the general performance level of the other systems. However the ICDSM achieves this performance only with a significantly lower number of parameters.

There are other interesting aspects of the ICDSM which are worthwhile to mention. One such aspect is that the ICDSM operates directly on the waveform space, something which makes it suitable for noise robust ASR. Another interesting point is that as the ICDSM can model the speech signal with a smaller number parameters, it is a good candidate for high performance speaker adaptation. We have experimented with those aspects and plan to report the results in a separate paper.

Model	approach	NFP	testset	%Accu
This paper	ICDSM alone	8k	core	64.11
	ICDSM+HMM	16k	core	67.86
Young [9]	triphone HMM	800k	full	72.3
Robinson [10]	recurrent net	47k	core	73.9
Chen [11]	HMM+context	N/A	random	70.4
Sun [12]	interpolation	25k	random	72.5

Table 4. *Comparison with the systems which handle co-articulation explicitly. Here NFP stands for number of free parameters.*

7. REFERENCES

[1] Don Hush, "Nonlinear signal processing methods," *IEEE Signal Processing Mag.*, vol. 15, no. 3, pp. 20–22, 1998.

[2] Herve Boulard, Hynek Hermansky, and Nelson Morgan, "Towards increasing speech recognition error rates," *Speech Communication*, vol. 18, pp. 205–255, 1996.

[3] Arun Kumar and S.K. Mullick, "Non-linear dynamical analysis of speech," *J. of the Acoustical Society of America*, vol. 100, no. 1, pp. 615–629, July 1996.

[4] F. Takens, "Detecting strange attractors in turbulence," in *Dynamical systems and turbulence*, D. Rand and L. S. Young, Eds., pp. 366–381. Springer Verlag Inc, 1981.

[5] Henry Abarbanel, *Analysis of observed chaotic data*, Springer Verlag Inc, 1996.

[6] Andrew Fraser, "Reconstructing attractors from scaler time series: a comparison of singular system and redundancy criteria," *Physica D*, vol. 34, pp. 391–404, 1989.

[7] Finn Tore Johansen, *Global discriminative modeling for automatic speech recognition*, Ph.D. thesis, University of Trondheim, Norwegian Institute of Technology, 1996.

[8] Narada D. Warakagoda and Magne H. Johnsen, "Neural network based optimal feature extraction for ASR," in *Proc. Conf. on Speech Communication and Technology(EUROSPEECH)*, 1999, vol. 1, pp. 97–100.

[9] S. J. Young and P. C. Woodland, "State clustering in hidden Markov model based continuous speech recognition," *Computer Speech and Language*, vol. 8, pp. 369–383, 1994.

[10] A. J. Robinson, "An application of recurrent nets to phone probability estimation," *IEEE Trans. Neural Networks*, vol. 5, no. 2, pp. 298–305, 1994.

[11] Ruxin Chen and Leah Jamieson, "Explicit modeling of coarticulation in a statistical speech recognizer," in *Proc. Int. Conf. on Acoustics, Speech, and Signal Proc. (ICASSP)*, 1996, vol. 1, pp. 463–466.

[12] D. X. Sun, "Statistical modeling of coarticulation in continuous speech based on data driven interpolation," in *Proc. Int. Conf. on Acoustics, Speech, and Signal Proc. (ICASSP)*, 1997, vol. 3, pp. 1751–1754.