Speaker Adaptation in an ASR System based on Nonlinear Dynamical Systems

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, which is solved using the ideas borrowed from the theory of Embedding. The resulting architecture, which we call Inverted CDSM (ICDSM) is well suited for modeling variations of speaker and channel characteristics, by its nature. We have evaluated the ICDSM using a set of experiments involving speaker adaptation in a continuous speech recognition task on the TIMIT database. Results of these experiments confirm the feasibility and potential advantages of the approach.

1. Introduction

From the statistical pattern recognition point of view, an ASR system contains two main elements: a feature extractor and a classifier. In mainstream approaches to ASR, the feature extractor is usually a Fourier spectral analyzer of one or another form. Further, the classifier is usually based on Hidden Markov Models which make use of probability distributions estimated using the assumption of uncorrelated successive feature vectors. However, both Fourier spectral analysis and neglect of feature vector correlations represent a highly simplified view of complex nonlinear phenomena involved in speech communication. A more realistic modeling approach, therefore, should allow us to drop the linearity assumption which is the foundation of Fourier analysis and to extract the correlations among successive feature vectors in an efficient manner.

Theory of nonlinear dynamical systems which considers the continuity of a physical process in nonlinear terms provides a solution to the above two problems simultaneously. In order to make use of this theory, we view the underlying mechanisms of a speech generation as a time varying nonlinear dynamical system. The time varying nature can be achieved by combining two "stationary" dynamical systems. We have considered a chain-like arrangement of dynamical systems for our work and we refer to this as the Chained Dynamical System Model (CDSM) [1]. 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 observed at the output. Therefore the ASR problem can be viewed as inverting the CDSM. In [1], we have proposed a solution to this inversion problem, which is rooted in Takens' embedding theorem [2]. The resulting architecture called Inverted Chained Dynamical System (ICDSM) can be used for practical speech recognition as reported in [1].

The ICDSM architecture, by its nature, is well suited for modeling speaker and environment variations as applied to ASR. One reason for this suitability is the possibility to transfer variation effects from one component of the ICDSM to another. Further, the ICDSM has an in-built noise filtering property, which makes it a natural candidate for environment/channel variation modeling. It can also be used for compensating for effects due to speaking rate variations. In this paper, however, we consider only the speaker variations, and show how the ICDSM can cope with those to enhance its robustness when used as a speech recognizer.

The rest of the paper is organized as follows. In section 2 we outline the structure of the CDSM and how the ICDSM can be used as a speech recognizer. Section 3 is devoted to a brief introduction to the variation transferability property of the ICDSM. Next, in section 4 some experiments carried out to evaluate speaker adaptation using the ICDSM and their results are presented. Finally, in section 5, we make some concluding remarks on the work.

2. CDSM and ICDSM

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 $\boldsymbol{\Phi}_2$. 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 $\boldsymbol{\Phi}_1$.

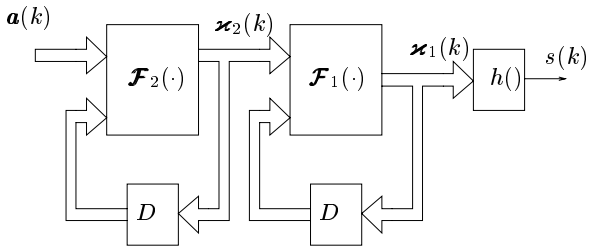

Figure 1: *Chained Dynamical System model (CDSM). Here D denotes a delay element.*

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 based on a generalized version of Takens' theorem [2] is shown in figure 2. In this figure, k_D and K_f are integer valued system parameters while $\mathbf{F}_p$ and $\mathbf{F}_2$ are (phoneme class independent) nonlinear functions implemented using Multilayer Perceptrons (MLPs). Effects which can not be described by the dynamical systems (including modeling errors) are represented by the component $\mathbf{Pr}_j$ for each class j, whose output is given by

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

where D_E is the dimension of the vector $\mathbf{e}(k)$ and $\mathbf{\Sigma}_j$ is its covariance matrix. Finally, note that $\mathbf{x}_2(k)$ is an estimate for the state vector $\boldsymbol{\varkappa}_2(k)$ and $\mathbf{a}(k)$ are the estimated control vectors which represent the phoneme class k^{th} speech vector belongs to. In other words, $\mathbf{a}(k)$, for a given k, can take a value from the set $\{\mathbf{a}_j \mid j = 0, 1, \ldots, C-1\}$ where C is the number of phoneme classes. See [1] for more details.

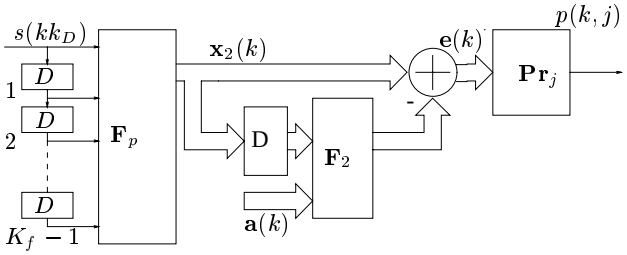

Figure 2: *The ICDSM suitable for practical ASR*

We can run a training procedure to estimate the parameters of the function blocks $\mathbf{F}_p$ and $\mathbf{F}_2$, as well as the covariance matrix $\mathbf{\Sigma}_j$ and the control input $\mathbf{a}_j$ for each class $j = 0, 1, \ldots, C-1$. We have opted to employ a gradient based training procedure, where an objective function defined as a function of $p(k,j)$ for all k and j over the training set is used. Since we can view the arrangement of $p(k,j)$ over k and j as a trellis (with $p(k,j)$ as its nodes), there is a straightforward similarity to the HMM based systems. Therefore standard algorithms found in the literature [3] can be used to compute the gradients of the objective function with respect to $p(k,j)$, and these can be back-propagated through the structure to obtain the gradients with respect to the parameters. We use two types of training algorithms; an isolated mode algorithm based on Maximum Likelihood (ML) and a discriminative continuous mode algorithm based on Maximum Mutual Information (MMI) criterion [3]. In both cases, a gradient technique based on the RPROP update rule is employed [3].

A procedure based on straightforward Viterbi search through the $p(k,j)$ trellis is employed to perform recognition.

3. Transferring variations within the ICDSM

Let us consider two different sets of conditions under which speech recognition has to be performed. Due to these differences we, in general, need two different ICDSMs to model the signal in those two cases. Let us assume that these ICDSMs have parameter sets $\mathbf{\Theta}_1$ and $\mathbf{\Theta}_2$ corresponding to

- $\{\mathbf{F}_p^1(\cdot), \mathbf{F}_2^1(\cdot), \mathbf{a}^1(k)\}$ and
- $\{\mathbf{F}_p^2(\cdot), \mathbf{F}_2^2(\cdot), \mathbf{a}^2(k)\}$,

respectively. Note that the differences in the two cases affect all three components in the set, making them different in the two cases. But it might be possible to keep, for example, one component constant through the two cases, by altering other components accordingly. This is what we mean by transferring variations within the ICDSM. Let us investigate what conditions must be satisfied to achieve this.

For example, consider the case where we would like to transfer the variations away from $\mathbf{F}_p$, i.e. to transform $\mathbf{F}_p^2$ to $\mathbf{F}_p^1$. Solution to this problem lies in the following conjecture.

Conjecture 3.1 *Let $\mathbf{G} : \mathbf{R}^D \to \mathbf{R}^D$ be a mapping, where D is the dimensionality of the state vector, such that $\mathbf{G} \circ \mathbf{F}_p^2 = \mathbf{F}_p^1$. If we can find such a $\mathbf{G}$ and some $\mathbf{F}_2^3 : \mathbf{R}^D \to \mathbf{R}^D$ such that $\mathbf{G} \circ \mathbf{F}_2^2 \circ \mathbf{g}_2^{\mathbf{a}^2} = \mathbf{F}_2^3 \circ \mathbf{g}_2^{\mathbf{a}^2} \circ \mathbf{G}$, then under ideal conditions (i.e. $\mathbf{e}(k) = \mathbf{0}$) the second ICDSM above can be represented by the parameter set $\{\mathbf{F}_p^1(\cdot), \mathbf{F}_2^3(\cdot), \mathbf{a}^2(k)\}$. In these expressions $\mathbf{g}_2^{\mathbf{a}^2}$ represents the auxiliary function which combines the delayed state vector $\mathbf{x}_2(k)$ and the control vector $\mathbf{a}^2(k)$.*

This conjecture can easily be verified by introducing a simple change of variables in the equation set defining the second ICDSM.

If the conditions stated in the conjecture are satisfied, then we manage to keep the same $\mathbf{F}_p^1(\cdot)$ for both cases, while lumping all the variations in the function block $\mathbf{F}_2(\cdot)$ and $\mathbf{a}$. Note however that we cannot make any statement as to whether those conditions are satisfied.

As a second example, we can demonstrate how all the variations are lumped into the function block $\mathbf{F}_p(\cdot)$. A conjecture which leads to the solution for this case is as follows.

Conjecture 3.2 *If we can find some function $\mathbf{G} : \mathbf{R}^D \to \mathbf{R}^D$ such that $\mathbf{G} \circ \mathbf{F}_2^1 \circ \mathbf{g}_2^{\mathbf{a}^1} = \mathbf{F}_2^2 \circ \mathbf{g}_2^{\mathbf{a}^2} \circ \mathbf{G}$, then the second ICDSM can be represented by $\{\mathbf{F}_p^3(\cdot), \mathbf{F}_2^1(\cdot), \mathbf{a}^1(k)\}$, with $\mathbf{F}_p^3(\cdot) = \mathbf{G} \circ \mathbf{F}_p^2$.*

This conjecture can also be verified easily as in the previous case. Note however that when compared to the previous conjecture, the conditions required to be satisfied here are less restrictive.

4. Speaker Adaptation

All the components included in the ICDSM, namely $\mathbf{F}_p(\cdot)$, $\mathbf{F}_2(\cdot)$ and $\mathbf{a}(k)$, are in general speaker dependent. This implies that each of these components has a role to play in speaker variation compensation. But as pointed out in the previous section, under some circumstances, one component can be made free from this responsibility, by transferring it to another.

In the context of state-of-the-art ASR, there are two main strategies for speaker adaptation; feature vector transform and model parameter transform [4]. However, for the case of ICDSM there is no clear distinction between these two strategies, because the analogous component to feature extractor, $\mathbf{F}_p$, is integrated into the trainable model. What is meaningful in our case is which part of the model ($\mathbf{F}_p(\cdot)$, $\mathbf{F}_2(\cdot)$ or $\mathbf{a}_j,\ j = 0, 1, \ldots, C-1$) is adapted.

To perform adaptation, we follow the approach in [5] , which is simply to run the gradient based training algorithm on the adaptation data using the SI model as the starting point.

This approach is not very sophisticated, but the simplicity is the motivation for this choice.

In our study we consider a supervised offline adaptation procedure as well as an unsupervised online procedure. In order to evaluate the ICDSM in these studies, we select the so called 39-class phoneme recognition task on the TIMIT database [3].

4.1. Speaker independent model

The system used as the speaker independent (SI) model for these experiments is prepared as follows [1]. Our starting point is the basic architecture depicted in figure 2. 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 section 2 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. The function block $\mathbf{F}_p$ is implemented as 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.

The dataset for training the SI model is prepared by picking all the SI and SX sentences (3696 sentences all together) from all 462 speakers in the TIMIT training set. Before training, any element which does not have a systematic method for initialization, is initialized randomly. Then the isolated mode training procedure, which is referred to as *phase 1* of training, is run. Finally, discriminative training is carried out in the continuous mode and we call it *phase 2* of training.

4.2. Supervised offline adaptation

We aim at performing adaptation on all three components, $\mathbf{F}_p(\cdot)$,$\mathbf{F}_2(\cdot)$ and $\mathbf{a}_j,\ \ j = 0, 1, \ldots, C-1$. However the function block $\mathbf{F}_p(\cdot)$ involves a large number of parameters and hence to achieve a meaningful adaptation one needs a sufficiently large amount of data, which will typically amount to several minutes [6]. But our database, TIMIT, contains only 10 utterances by each speaker, and therefore it does not provide sufficient data to perform true speaker adaptation of the above kind. Due to this reason, in this subsection, we decide to perform *sex adaptation* instead of adaptation to each individual speaker. This means that we try to build separate male and female models starting from our speaker independent models.

The required data set to run offline adaptation is created in the following way. 100 sentences, each uttered by a female, are randomly drawn from the TIMIT complete test set. But it is maintained that none of these sentences are included in the TIMIT core test set, and all 8 dialect regions are covered. This set of sentences is referred to as *female adaptation set*. The same is done for males and the resulting set is called *male adaptation set*. After [5], each of these sets are divided into two subsets, a training set (for adaptation) of 80 sentences and a cross validation set of 20 sentences.

To obtain the female-sex dependent model, phase 2 of the usual training procedure mentioned in section 4.1 is run on the female adaptation set starting from the SI model. Note however that only the parameters in the component set $\mathcal{S}$ are updated during this procedure , where $\mathcal{S}$ contains either $\mathbf{F}_p(\cdot)$,$\mathbf{F}_2(\cdot)$, $\mathbf{a}_j$, or a combination of them depending on the experiment. Cross validation is used to monitor the progress of this training procedure. The resulting model set is then tested on all female utterances in the TIMIT core test set. The same procedure is repeated for the male adaptation set and tested on the male utterances in the core test set. Recognition scores are generated for the union of male and female test results, which are shown in Table 1.

Adapted component	%Corr	%Accu
none	65.39	61.62
$\mathbf{F}_p(\cdot)$	68.28	64.62
$\mathbf{F}_2(\cdot)$	67.46	62.52
$\mathbf{a}$	67.61	62.50
$\mathbf{a}$ & $\mathbf{F}_2(\cdot)$	67.84	63.63
$\mathbf{a}$,$\mathbf{F}_2(\cdot)$ &$\mathbf{F}_p(\cdot)$	68.49	64.85

Table 1: Recognition results for adaptation of different components of the ICDSM, in a supervised sex adaptation task

Figures in the table 1 clearly shows that the adaptation procedure described leads to improvements in the recognition results. Also noted is that adaptation of $\mathbf{F}_p(\cdot)$ is more effective than adapting the other two components either individually or simultaneously. Another observation is that adapting $\mathbf{F}_p(\cdot)$ alone and adapting all three components, result in almost the same level of performance. One possible reason for this behaviour is the fact that the requirement for transferring variabilities from $\mathbf{F}_p(\cdot)$ to the other components is more restrictive than for doing it other way around.

4.3. Unsupervised online adaptation

Unlike the supervised, offline adaptation which uses a relatively large amount of data, unsupervised adaptation in online fashon usually operates under the restriction that only a very small amount of data is available. Therefore, online adaptation can be expected to perform satisfactorily only when the adapted parameter set is sufficiently small. Due to this fact, the function block $\mathbf{F}_p(\cdot)$ which has a large number of parameters disqualifies to be used as an adaptive component in these experiments. However other two components $\mathbf{F}_2(\cdot)$ and $\mathbf{a}_j,\ j = 0, 1, \ldots, C-1$ contain only a small numbers of parameters, and hence they are considered for adaptation in the experiments described below.

Since online adaptation is run on a small amount of data (typically a few seconds of speech), number of utterences per speaker provided in the TIMIT database is sufficient for "true speaker adaptation" experiments. Therefore we do not have to resort to adapation procedures which considers broader classes of variations such as sex-adaptation. Further, we do not need a seperate adaptation database, because adaptation is performed on the test set itself.

As in the previous subsection, starting point of the experiments is the speaker independent model described in subsection 4.1. The adaptation procedure, which is run directly on the test set for each speaker, consists of the following two steps.

1. Present a set of utterences (say, consisting of N_A utterences) to the system, and run the usual recognition pro-

cedure. This will produce a transcription for each utterence

2. Using the transcription obtained in step 1, run N_I iterations of (only phase 2 of) the usual training procedure sketched in subsection 4.1 on those N_A sentences. But update only the parameters in the component set of interest $\mathcal{S}$, where $\mathcal{S}$ can include $\mathbf{F}_2(\cdot)$, $\mathbf{a}_j$, $j = 0, 1, \ldots, C-1$, or both. Other parameters of the system are frozen at the speaker independent values.

These two steps are repeated until the transcriptions produced in step 1 do not change. Final recognition scores are produced on these transcritions collected over all speakers in the test set. In our experiments, adaptation batch size N_A is selected to be 1, 4 and 8. The number of training iterations during adaptation N_I is set to 10. Results for these experiments are shown in table 2.

Adapted component	N_A	%Corr	%Accu
none	-	65.39	61.62
$\mathbf{F}_2(\cdot)$	1	64.92	60.63
	4	66.81	62.20
	8	67.74	63.25
$\mathbf{a}$	1	65.18	60.82
	4	67.17	62.28
	8	67.78	63.32
$\mathbf{F}_2(\cdot)$ and $\mathbf{a}$	1	64.13	61.08
	4	67.70	62.87
	8	68.11	64.08

Table 2: Recognition results for unsupervised speaker adaptation of different components of ICDSM

From table 2, it is clear that unsupervised adapatation can lead to better recognition results. Adaptation of $\mathbf{F}_2(\cdot)$ or $\mathbf{a}$ does not seem to have any sigificant advantage over the other. However, adaptation of both components ($\mathbf{F}_2(\cdot)$ and $\mathbf{a}$) gives slightly better recognition rates than adaptation of a single component. This is not a surprise, because with a higher number of parameters, it is easier to absorb the variations. This causes a recognition improvement as long as the generalization ability is not affected. Another point which is worthwhile to notice is that when the adaptation batch size (N_A) is 1, recognition rates decrease wrt the unadapted case. This may have something to do with the wrong transcriptions generated in the step 1 of the online adaptation procedure. If we have only 1 sentence in the batch, then there is no mechamism to correct such an incorrect transcription subsequently[1]. The converse of this fact can be used to explain the observation that as the batch size N_A increases the recognition rate improves.

5. Concluding remarks

In this paper we tried to minimize the effects of speaker variations on an ICDSM based ASR system through adapatation. We showed that, even though speaker variations affect all components of the ICDSM, under certain conditions these effects can be transferred to other components. This makes it possible, in principle, to compensate for all the variabilities by manupulating only a single convenient component. But in practice, these transferring possibilities become limited partly because of the difficulties in satisfying the required conditions. Intuitive examination of these conditions suggests that transferring variabilities away from $\mathbf{F}_p$ is more restrictive than doing it in the other way around. In fact, the results of our experiments support this proposition. This is however somewhat unfortunate, because $\mathbf{F}_p$ is the largest in terms of the number of parameters and hence the one which imposes the tightest requirements on adaptation data and other resources. Nevertheless, adaptation of low parameter components ($\mathbf{F}_2$ and $\mathbf{a}$) leads also to comparable performance improvements, especially in the case of on-line adaptation.

Even though adaptation improves the performnce of the ICDSM, one may notice that general performance level of the system reported here is low, compared to the best systems such as [7] applied to the same task. As reported in [1], we have tried to improve the performance level by combining the ICDSM with a HMM based system. This actually works, for the unadapted case, but when the ICDSM is adapted, the overall performnace becomes poorer, possibly due to the mismatches created by the adaptation process. Another approach we have tried to improve the recognition accuracies is to use a mixture of $\mathbf{F}_2$ blocks (instead of a single block). This too leads to improved recognition rates, but it is yet to be used for adaptation experiments.

Finally, it is worthwhile to mention that there are other interesting properties of the ICDSM which can be used for variation compensation. 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 the possibility to include the speaking rate invariance to the system in a natural way. We have experimented with those aspects and the results will be reported elsewhere.

[1] In addition to by-chance corrections associated with the parameter update procedure

6. References

[1] Narada Warakagoda and Magne H. Johnsen, "Nonlinear dynamical system based acoustic modeling for ASR," in *Proc. Int. Conf. on Acoustics, Speech, and Signal Proc. (ICASSP)*, 2001, Accepted to be published.

[2] 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.

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

[4] A. Sankar and C. H. Lee, "A maximum likelihood approach to stochastic matching for robust speech recognition," *IEEE Trans. on Speech and Audio Processing*, vol. 4, no. 3, pp. 190–202, 1996.

[5] J. P. Neto, G. Martins, and L. Almeida, "Speker adaptation in a hybrid HMM-MLP recognizer," in *Proc. Int. Conf. on Acoustics, Speech, and Signal Proc. (ICASSP)*, 1996, vol. 6, pp. 3382–3385.

[6] S. Young, Joop Jansen, Julian Odell, Dave Ollason, and Phil Woodland, *The HTK Book*, Entropic Research laboratories, 1996.

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