

Gaussian Dynamic Warping (GDW) Method applied to Text-Dependent Speaker Detection and Verification

Jean-François Bonastre^(1,2), Philippe Morin⁽²⁾, Jean-Claude Junqua⁽²⁾

(1) LIA, University of Avignon, Avignon, France

(2) Panasonic Speech Technology Laboratory, Santa Barbara, USA

(jfb@lia.univ-avignon.fr, phm@research.panasonic.com, jcj@research.panasonic.com)

Abstract

This paper introduces a new acoustic modeling method called Gaussian Dynamic Warping (GDW). It is targeting real world applications such as voice-based entrance door security systems, the example presented in this paper. The proposed approach uses a hierarchical statistical framework with three levels of specialization for the acoustic modeling. The highest level of specialization is in addition responsible for the modeling of the temporal constraints via a specific Temporal Structure Information (TSI) component. The preliminary results show the ability of the GDW method to elegantly take into account the acoustic variability of speech while capturing important temporal constraints.

1. Introduction

Two main classes of techniques based on either DTW or HMM [3][4][9] are typically used in isolated word recognition or text-dependent speaker recognition. Dynamic Time Warping (DTW) systems are well suited when the amount of enrollment data is small and have the ability to precisely model time constraints. In addition, they can easily be adapted to support the spotting of target events at a reduced computational cost. A target event can represent a word or phrase alone or a word or phrase along with the speaker's identity. The DTW approach lacks however the generalization power available with HMM and GMM (Gaussian Mixture Models) approaches to deal with the variability inherent to the speaker or to the environment. Another drawback is also their inefficiency at taking advantage of larger amount of training or adaptation data.

In contrast, HMM-based techniques allow for a good estimation of the target events' acoustic space and provide a solid framework for score normalization and model adaptation. However, HMM modeling requires a large amount of training data and can be expensive in terms of resources. GMMs [2], which are single-state HMMs, address some of these issues and achieve reasonable performance especially in the case of text-independent speaker recognition with lower requirements in training material. However the main disadvantage of GMM systems and – to a lesser extent – HMM systems resides in their inefficiency at

preserving and modeling the temporal aspects of speech (TSI), strongly limiting their performance in the case of short duration phrases especially when word-spotting is a desired feature.

Solutions for combining statistical modeling and TSI constraints have been proposed. In [5], the TSI aspects are embedded in the GMM components. In [6], the TSI is taken into account by a trajectory model.

In order to model the TSI of speech while achieving high modeling performance, low cost decoding and enabling an efficient spotting mode, we propose a new target event modeling approach that is based on both statistical acoustic space modeling and TSI constraints modeling. The proposed approach [1] called Gaussian Dynamic Warping (GDW) provides a highly flexible framework that combines together the advantages found in both HMM/GMM-based and DTW-based systems.

This paper presents the GDW approach (see Section 2) applied to a speaker detection and verification task. Section 3 shows preliminary experiments realized with PSTL's voice entrance door system and Section 4 summarizes the benefit of the approach and presents future work directions.

2. The GDW approach

The GDW approach mixes statistical modeling and Temporal Structure Information (TSI) modeling. A GDW model is a hierarchical statistical model with three specialization levels. All the components of the GDW model are statistically trained. Each statistical node can be a GMM or an HMM component. Those nodes (as explained later) can share means, variances, weights and topology information.

The TSI constraints are used by a DTW-based module that can either perform endpoint-based matching during the training phase or perform phrase-spotting during recognition/verification phase.

The core subsystem proposes the high level functionality, like enrolment and recognition.

2.1. GDW Hierarchical Statistical Modeling

Statistical modeling is at the core of the GDW approach. It is based on statistical modeling with an original three levels hierarchical structure shown Figure 1.

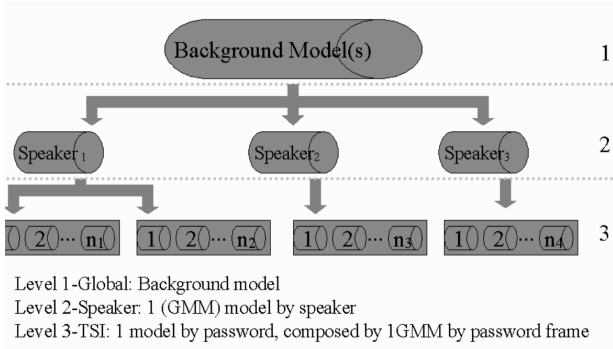


Figure 1: Overview of the GDW model

In the current experiments, GMM modeling was chosen and each GMM statistical node have the same dimension in terms of the number of Gaussian components.

The **top layer** is the layer with the least specialization and is represented as a classical Background Model denoted BM that is trained using EM/ML [7].

The **middle layer** captures the speaker-specific spaces. Each speaker-specific model (denoted X) is derived from the background model BM using a MAP approach [8]. All the acoustic data available for a given speaker is used during this phase. It represents an important difference when compared to the classical approach for which the adaptation is done independently phrase by phrase. In effect, only a limited number of Gaussians need to be adapted from the background model and mean adaptation alone can be performed.

The **bottom level** captures for each speaker the phrase-specific time-dependent sub-spaces. A TSI frame-dependent statistical node is trained/derived for each time frame (e.g. 10 ms speech segments) of the word or phrase using a specific MAP algorithm. In order to reduce the computational cost, it was found that weight adaptation of the N-best Gaussians alone is sufficient to provide good performance.

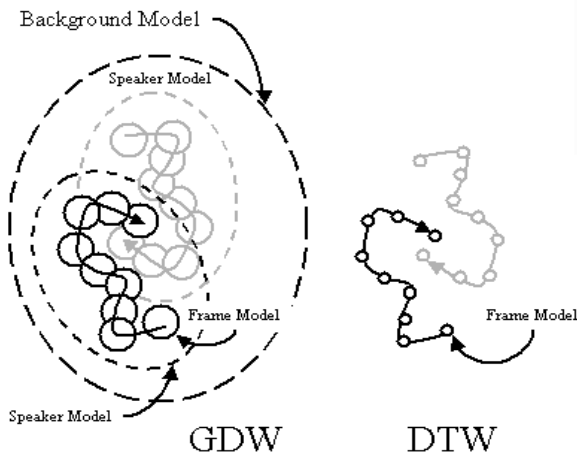


Figure 2: Comparison between GDW and DTW modeling

Figure 2 illustrates the basic difference between a DTW approach and a GDW approach. It shows that a DTW approach essentially lacks the generalization and

comparative power brought by the GDW approach via its layered statistical modeling since phrase models are built and recognized independently of each other. Regarding the GDW approach, it also shows that the degree of specialization attached to a frame-dependent model is not fixed but is on the contrary highly adjustable, from a single frame-independent model equivalent (i.e. a same GMM for all the password frames) up to a hyper-specialized model (i.e. DTW-like state).

2.2. Temporal Structure Information Processing

In the GDW approach, the Temporal Structure Information (TSI) is represented as a set of connected frame-dependent states. Each state represents a piece of information that is both specific to the phrase and to the speaker at this instant in time. Each TSI state is currently modeled as a GMM structure and is derived from the corresponding level-two model (the speaker model). The degree of specialization of TSI states can be adjusted at training time. This frame-dependent state adaptation scheme enables a smoothing and optimization of the representation based both on the quantity and quality of training data available and on the ratio between local information and a priori ratio between generalization and specialization. Depending of this ratio, the state models are moved from the speaker model (all the state models are equal) to a frame-based representation (a state model corresponds mainly to the current frame).

The Temporal Structure Information is preliminary used for detecting the target event in the incoming audio stream. A Dynamic Time Warping-based algorithm is used to spot the target events in real-time. The spotting module generates a list of hypotheses (possibly an empty list) at each time instant. Each hypothesis (described by a detection score and an alignment path) is then checked by the verification module. The detection module uses a set of alignment penalties for insertions and deletions.

$$LocalDist(y, X_n) = NormDist\left(\log\left(\frac{l(y|X_n)}{l(y|X)}\right)\right)$$

Equation 1: y is the input frame, X_n is the frame model, X the global event model. $NormDist()$ is a normalization function used to transform a log likelihood ratio into a $[0,1]$ "distance".

The DTW algorithm uses a statistical local distance to measure the degree of similarity between an input frame and a frame-dependent GMM. The local distance is derived from a log likelihood ratio. The numerator of the ratio is the likelihood of the frame given the TSI frame-dependent model, denoted X_n and the denominator is the likelihood of the frame given the speaker model X (Equation 1).

The detection score computed by TSI spotting module is the sum of the local distances augmented by the DTW-like penalties along the path normalized by the number of input frames in the path.

2.3. Training of a GDW Target Event Model

The training of a GDW target event model requires the training of a speaker model (denoted X) and the training of TSI state-dependent models. The TSI state-dependent models are trained by adaptation of the X model. Figure 3 gives an overview of the training process.

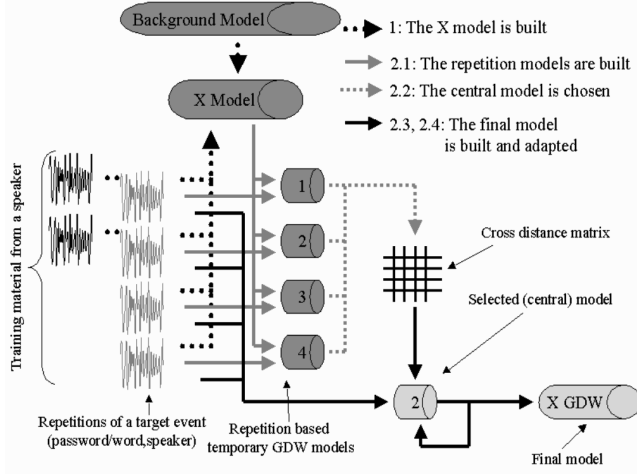


Figure 3: Overview of GDW model training process

The training of the speaker model is accomplished by using all the training speech data available for that speaker, thus it is not limited at using only the speech data corresponding to the phrase to be modeled.

The training of the TSI state-dependent models is done as follows:

- First, a GDW model is built independently for each enrollment repetition available. In this case, the number of TSI state-dependent models correspond to the number of frames and each state-model is set using only the corresponding frame.
- Secondly, the cross-matching distance matrix between all the enrollment repetitions of the phrase is computed. The phrase providing the minimum average distortion is selected as “central” repetition and is used as seed for the next step.
- Thirdly, the enrollment repetitions are aligned with the central model and the central model is adapted. This step is iterated several times until a convergence criterion (based on the minimum total distortion) has been reached.

The adaptation of a TSI state model is done using all the frames (from the different enrolment utterances) aligned with it. In order to save memory and computational resource, only the weight of a subset of the components are adapted using a specific MAP interpolation function defined by equation 2. The components with the larger weights are selected.

$$w_i^{X_n} = \frac{w_i^X + (1 - \alpha) \hat{w}_i^{X_n}}{1 + (1 - \alpha)}$$

Equation 2: $w_i^{X_n}$ is the final (adapted) weight of the i component, $\hat{w}_i^{X_n}$ is the weight computed using the adaptation data, w_i^X the a priori weight (from X) and α , the adaptation factor.

2.4. Target Event Detection and Verification

The recognition of a target event is achieved as a two-stage process:

- Firstly, as explained earlier, the spotting module generates a list of possible speaker-phrase hypotheses at each time instant. The hypotheses having a detection score that is better than the detection threshold are passed to the next stage.
- Secondly, a speaker verification score is computed and compared to a decision threshold.

The speaker verification score is obtained as the average over all the frames in the spotted path of the local biometric score called *BioScore*, given by equation 3. It is close to the traditional log likelihood ratio used in speaker recognition.

$$BioScore(y, X_n) = \left(\frac{local \cdot \log(l(y|X_n)) + (1 - local) \cdot \log(l(y|X))}{(1 - local) \cdot \log(l(y|BM))} \right) - \log(l(y|BM))$$

Equation 3: y denotes the input frame, X_n the frame-based speaker model, X the speaker model, BM the background model and $local$, a weight used for local/global tuning.

The local biometric score measures the degree of similarity between an input frame and the target event’s aligned GMM component. It is normalized by the BM model, in order to minimize the influence of non-informative frames (non-speech frames for example) and noisy frames. The weight of the frame-dependent TSI model compared to the global speaker model is given by the *local* parameter. Usually, the *local* parameter is set close to 1 in order to give a greater control to the frame-dependent models.

3. Preliminary Experiments

We implemented and tested the GDW approach with the speech data collected by a proprietary vocal entrance door system developed and installed at PSTL. The data set correspond to natural spontaneous speech collected over several weeks. With that system, each user had enrolled the password of his/her choice. Most passwords were fairly short and ranged from 0.6 to 1.4 seconds in duration with a 0.8 second average. In this application, enrolled users could simply come near the biometric box and say their password to unlock the door. With that system, no control over the environmental noise (it is an outside door) was made and speakers were found to address the system from a variable distance ranging from about 7 inches up to 10 feet.

3.1. Data Set and Experimental Protocol

The preliminary experiments were conducted using a PSTL's proprietary data set described above. That database contains speech data collected at 8KHz for a total of 27 users. At least five repetitions of a user-selected password were used for enrollment. The test set was composed of 242 client tests and 311 impostor tests. For all the impostors tests, knowledge of the correct password by the impostor was assumed.

A third data set was dedicated to the training of the background model. The vocabulary used during the third set recordings includes the client passwords and some speech from the enrolled speakers were included.

3.2. Results

The number of components in the GMM background model (the top level of GDW model hierarchy) was fixed to 512 components. The speaker models (second layer of the GDW approach) were derived from the background model by adapting 256 components. The TSI frame models were adapted by changing only 16 component weights.

local/global ratio	BM weight	HTER min (%)
0	0.1	5.4
0.5	0.1	3.5
1	0.01	3.1
1	0.1	2.8
1	0.5	2.8
1	0.75	3.2

Table 1: HTER obtained with the GDW method, based on different values for the contribution ratio between local and global models in bioscore computation and the weight of BM model used during local (TSI) models MAP adaptation.

The table 1 presents a synthetic view of a small set of experiments. It shows clearly that the TSI constraints used by GDW method contribute to a significant reduction of the errors (from 5.4% to 2.8% of half total error rate) when the biometric score computation uses local TSI models (with local=1) as opposed to a global speaker model (with local=0).

This table shows also that the tuning of TSI models weight adaptation factor (alpha) is not critical for GDW. It shows that – thanks to the specific MAP formula - the GDW models converge for all the ratio specialization/globalization.

4. Conclusions and Future Work

In this paper, we presented a new acoustic modeling method called Gaussian Dynamic Warping (GDW). Its main originality is to elegantly combine the power of statistical modeling with Temporal Structure Information (TSI) constraints modeling traditionally used in DTW-based systems. The GDW approach can efficiently model target events with a limited amount of training data. The GDW modeling technique provides a

flexible and tunable framework that can accommodate a variety of recognition and verification tasks.

This paper described the first implementation of GDW approach with an evaluation on a natural and spontaneous speech database. These preliminary experiments show the interesting potential of the approach. Further research directions will include an evaluation of the method on larger databases for both speaker verification and speech recognition tasks. Particularly, a comparison of GDW approach with classical DTW, GMM and HMM systems will be performed.

The adaptation of GDW models to new environments will also be explored, by moving only the upper level of the hierarchy.

5. Acknowledgements

This work was done at Panasonic Speech Technology Laboratory (PSTL) in collaboration with the first author during his stay as a visiting professor. The authors would like to thank the University of Avignon (France) and PSTL (USA) for their continued support.

6. References

- [1] J.-F. Bonastre, P. Morin, J.C. Junqua, Gaussian Model-Based Dynamic Time Warping System and Method for Speech Processing, US PATENT Proposal 9432-000216/US, 2003.
- [2] D. A. Reynolds, "Speaker identification and verification using Gaussian mixture speaker models," *Speech Communication*, vol. 17, pp. 91-108, August 1995.
- [3] F. Jelinek, Statistical methods for speech recognition. MIT Press, Cambridge MA, 1997
- [4] L. Rabiner and B.-H. Juang, Fundamentals of speech recognition. Prentice Hall, 1993.
- [5] I. Illina, M. Afify, Y. Gong. Environment Normalisation Training and Environment Adaptation using Mixture Stochastic Trajectory Model. *Speech Communication*, vol. 26, n. 4, 1998
- [6] R. Stapert, J. S. Mason, "A Segmental Mixture Model for Speaker Recognition", Proc. Eurospeech, volume 4, pages 2509-2512, 2001
- [7] A. Dempster, N. Laird, D. Rubin, "Maximum likelihood from incomplete data via the EM algorithm", *J. Roy. Stat. Soc.*, Vol. 39, pp. 1-38, 1977.
- [8] J. L. Gauvain, C. H. Lee, "Maximum A Posteriori estimation for multivariate gaussian mixture observation of markovs chains", *IEEE Transactions on Speech and Audio Processing*, Vol. 2(2), pp. 291-298, April 1994
- [9] A. Higgins, L. Bahler, and J. Porter, Speaker Verification using randomized Phrase Prompting », *Digital Signal Processing*, Vol 1, p. 89-106, 1991