Structure-Based and Template-Based Automatic Speech Recognition --- Comparing parametric and non-parametric approaches

Li Deng [1], *Helmer Strik* [2]

[1] Microsoft Research, One Microsoft Way, Redmond, WA, USA
[2] CLST, Department of Linguistics, Radboud University, Nijmegen, the Netherlands
deng@microsoft.com, strik@let.ru.nl

Abstract

This paper provides an introductory tutorial for the Interspeech07 special session on "Structure-Based and Template-Based Automatic Speech Recognition". The purpose of the special session is to bring together researchers who have special interest in novel techniques that are aimed at overcoming weaknesses of HMMs for acoustic modeling in speech recognition. Numerous such approaches have been taken over the past dozen years, which can be broadly classified into structured-based (parametric) and template-based (non-parametric) ones. In this paper, we will provide an overview of both approaches, focusing on the incorporation of long-range temporal dependencies of the speech features and phonetic detail in speech recognition algorithms. We will provide a high-level survey on major existing work and systems using these two types of "beyond-HMM" frameworks. The contributed papers in this special session will elaborate further on the related topics.

Index Terms: structure-based, template-based, automatic speech recognition

1. Introduction

While hidden Markov modeling (HMM) has been the dominant technology for acoustic modeling in automatic speech recognition for a few decades, many of its weaknesses have also been well known and they have become the focus of much intensive research [e.g., 1, 2, 3, 5, 10, 11, 8, 23, 32, 37, 38, 41, 42]. One prominent weakness in current HMMs is the handicap in representing long-span temporal dependency in the acoustic feature sequence of speech, which, nevertheless, is an essential property of speech dynamics [e.g., 3, 7, 32]. The main cause of this handicap is the conditional IID (Independent and Identical Distribution) assumption inherit in the HMM formalism. In hidden Markov modeling it is also assumed that speech can be described as a sequence of discrete units, usually phonemes. In this symbolic, invariant approach the focus is on the verbal information, and the incoming speech signal is normalized during pre-processing in order to strip off most of the non-verbal (indexical) information. However, experiments have shown that this non-verbal information plays an important role in human speech recognition. Another weakness of (standard) HMMs is that it is difficult to include this indexical information [e.g., speaker-specific properties, fine phonetic detail) [37, 38]. Some of these difficulties have been addressed in a bottom-up, detection-based framework [23].

The purpose of this special session is to bring together researchers who have special interest in novel techniques that are aimed at overcoming weaknesses of HMMs for acoustic modeling in speech recognition. In particular, we plan to address issues related to the representation and exploitation of long-range temporal dependency in speech feature sequences, the incorporation of speaker variations and fine phonetic detail in speech recognition algorithms and systems, comparisons of pros and cons between the parametric and non-parametric approaches, and the computation resource requirements for the two approaches.

This paper is aimed to provide an overview of both structured-based (parametric) and template-based (non-parametric) approaches that have been developed in the past for overcoming the main weaknesses of HMMs. In Sections 2 and 3, we will introduce the structure-based and the template-based approaches, respectively. We will finish, in section 4, with discussion and conclusions.

2. Structure-based approach

The structure-based approach establishes mathematical models for stochastic trajectories or segments of speech utterances using various forms of parametric characterization. These parametric trajectory models include the use of piecewise polynomials [e.g., 5, 16, 32], linear dynamic systems [8, 12, 32], and nonlinear dynamic systems embedding hidden structure of speech dynamics [e.g., 6]. All these piecewise trajectory/dynamic models can be considered as generalizations of the HMM in relaxing the IID or piecewise constant assumptions.

Although HMM-based recognition systems perform well in many relatively simple speech recognition tasks, they do not model some important dynamic aspects of speech directly (and are known to perform poorly for difficult tasks such as conversational speech). As a result, they are not able to accommodate dynamic articulation differences between the speech signals used for training and the speech signal being decoded. For example, in casual speaking settings, speakers tend to hypo-articulate their speech. This means that the trajectory of the user's speech articulation may not reach its intended target before it is redirected to a next target. Because the training signals are typically formed using a "reading" style of speech in which the speaker provides more fully articulated speech material than in hypo-articulated speech, the hypo-articulated speech does not match the trained HMM states. As a result, the recognizer provides less than ideal recognition results for casual speech.

A similar problem occurs with hyper-articulated speech. In hyper-articulated speech, which often occurs in noisy environments, the speaker exerts extra effort to make the different sounds of their speech distinguishable. This extra effort can include changing the sounds of certain phonetic units so that they are more distinguishable from similar

sounding phonetic units, holding the sounds of certain phonetic units longer, or transitioning between sounds more abruptly so that each sound is perceived as being distinct from its neighbors. Each of these mechanisms makes it more difficult to recognize the speech using an HMM system because each technique results in a set of feature vectors for the speech signal that does not match well to the feature vectors present in the training data. HMM systems also have trouble dealing with changes in the rate at which people speak. Thus, if someone speaks slower or faster than in the training data, the HMM system will tend to make more errors decoding the speech signal.

Careful modeling of speech dynamics, especially when the dynamics are represented in a hidden or unobserved "articulatory" domain instead of in the observed acoustic domain, is capable of overcoming all of the above problems in the HMM. The essence of such a hidden-dynamic approach is that it exploits knowledge and mechanisms of human speech production so as to provide the structure of the multi-tiered stochastic process models. Most "hidden dynamic" models of speech use parametric forms to represent both the hidden dynamic vectors and the observed acoustic feature vectors (see a comprehensive survey of these models in [7]). These models are based on the underlying mechanisms of speech coarticulation and reduction, and on the relationship between speaking rate variations and the corresponding changes in the acoustic features. A specific layer in this type of models represents long-range temporal dependency of the hidden trajectory vectors in the form of (non-recursive) FIR filter, which is parameterized by exponentially decaying FIR filter coefficients in both forward (for anticipatory coarticulation) and backward (for carrying-over coarticulation) directions. Recent research using this type of hidden trajectory model in combination with HMMs has shown substantially lower phonetic recognition error rates compared with the state-of-the-art HMM system alone [e.g. 6, 42].

Another major type of hidden dynamic model uses recursive forms to parametrically represent the hidden speech dynamic vectors, and then uses a similar kind of nonlinear mapping to link the hidden vectors to the observed acoustic features. These models belong to a very wide class of switching nonlinear dynamic models, and have been developed by various groups of researchers [e.g., 9, 40] and surveyed in [7].

We would like to point out that although the main motivations of the structure-based parametric dynamic speech models are to parsimoniously embed the main-stream phonetic properties of speech such as coarticulation and phonetic reduction into the ASR algorithms, more detailed phonetic information such as systematic speaker variation can also be satisfactorily handled. The recent work by Yu et al. [42] developed a novel speaker-adaptive learning algorithm for the hidden trajectory model of [6]. The vocal tract resonance targets are key parameters of the model that control the hidden dynamic behavior and the subsequent acoustic properties. A speaker-adaptive training technique is reported that takes into account the variability in the target values among individual speakers. The adaptive learning is applied also to adjust each unknown test speaker's target values towards their true values.

As the parametric, structure-based speech dynamic models are further developed in the future, we expect that more fine phonetic detail will be incorporated within consistent statistical frameworks in a similar manner to the way speaker variation is handled as in [42].

3. Template-based approach

Differing from the structure-based approach which is developed using parametric models for data variation, the template-based approach relies directly on the observed training sequences based on non-parametric representations. We first provide motivations of the template-based approach for ASR from the perspectives of psycholinguistics and human speech recognition.

In almost all current HMM-based ASR systems, a rather similar paradigm is used in which utterances are represented as a sequence of words (language model), words as a sequence of phonemes (lexicon), and phonemes as a sequence of states (acoustic models) [4, 36, 37]. In this invariant, symbolic approach the focus is on recognizing words, the verbal information. This approach is challenged by recent psycholinguistic findings on the special roles of non-verbal information, fine phonetic detail, and phonetic variation. These findings are summarized here.

Speech contains two types of information: (1) verbal information and (2) non-verbal (indexical) information. Verbal information is mainly related to the content of the message, while indexical information is more related to the form, such as properties of the speaker (e.g. F0 and speech rate). With respect to indexical information, some interesting findings have been reported in the literature recently. Both indexical and non-indexical properties of speech appear to be stored by humans [15, 33]. Familiarity with a person's voice facilitates recognition of that person's speech [14, 15, 33], and facilitation also occurs for speakers whose speech is similar [14, 15]. Also for visual perception it has been found that familiar patterns are perceived better than unfamiliar ones [19, 20]. Besides these findings on indexical information, experimental results also show that fine phonetic detail can influence lexical access [17, 18, 28, 29, 35], such as, e.g., sub-phonemic differences between realizations of the monosyllabic word "ham" and the first syllable of "hamster" [35].

Because these findings on indexical information and fine phonetic detail are difficult to explain in current models of spoken word recognition, there is a growing belief that new models are needed, both in the field of psycholinguistics [see, e.g., 24] and ASR [see, e.g., 4, 30, 31, 34, 36, 37, 38]. Currently, there is an increasing interest in the template-based approach, which is also referred to as episodic, multiple trace, exemplar, example-based, or instance-based approach [1, 2, 10, 11, 13, 14, 15, 17, 21, 22, 25, 26, 27, 33, 37, 38].

What are the main differences between HMM-based and template-based ASR systems? In the HMM-based approach words are stored in the lexicon in the form of sequences of abstract (usually phonemic) symbols. Often, for some words more than one entry per word are present in the lexicon in order to model pronunciation variation, and these pronunciation variants are usually stored as different transcriptions in terms of phonemes [36, 39]. In the template-based approach words are not represented as sequences of symbols, but as sequences of (abstract) units that are represented in the form of many episodes (trajectories); a large number of episodes (traces) are stored, i.e. single instances of stimuli with many details, instead of the canonical representations stored in the invariant approach. In a template-based system the incoming signal is then compared to sequences of these stored episodes, e.g. sequences of feature vectors are compared. In an HMM-based system, the signal is compared to a sequence of states, and for each state

the conditional probability of a frame given that state is calculated by means of the stored probability density functions (pdf's, usually Gaussian mixtures) or the stored artificial neural networks. Probability density functions and artificial neural networks are parametric representations of all feature values observed in a large training corpus.

Template-based speech recognition requires large amounts of memory and computing power. Therefore, until recently, it has been practically impossible to investigate template-based approaches to speech recognition. Today, however, the computing power and memory that are needed to investigate the template-based approach to speech recognition are rapidly becoming available. Recently some research has started on using template-based approaches for ASR, and the initial results are promising [1, 2, 10, 11, 16, 25, 26, 27]. It has been shown that by combining HMM-based and template-based ASR the performance can be improved [1, 2, 11]. An important issue in template-based ASR is to find a suitable distance metric [2, 10, 26].

4. Discussion & conclusions

HMMs have been the dominant technology for a few decades in ASR. The performance of HMM-based ASR systems has gradually improved over the years. However, it is well known that HMMs have limitations. Two approaches to overcoming limitations of HMMs are described in the current paper: structure-based and template-based approaches. For both approaches, improvements in performance have been reported; however, these improvements were established in very different ways, using different techniques. Up until now, these two approaches have mainly developed independently of each other. In this special session we will try to bring them together. What are the advantages and disadvantages of these approaches? Can they be combined? In recent research, improvements of ASR performance have been obtained by combining (conventional) HMM and template-based techniques [1, 2, 11]. Similar combinations and performance improvements are also reported for the structure-based approaches [6, 41, 42]. Both groups of researchers have found that HMMs are rather robust, providing basic acoustic scores upon which the new approaches are acting and enhancing. A large part of the performance improvements is likely due to better modeling of long-range temporal dependencies, which HMM systems alone are not able to accomplish. Therefore, one could wonder how much additional improvement could be obtained by combining structure-based and template-based approaches, in addition to the combination with HMMs. Furthermore, it has been traditionally held that for more complex approaches and models it is more difficult to incorporate detailed knowledge into the algorithms. For example, in HMMs, speaker variation can be simply handled by pooling all data from many speakers in training. But for more complex hidden trajectory models, such pooling does not work since some key parameter set (i.e., resonance targets) in the model are speaker specific. Special normalization techniques are required [42]. How can other types of variations in speech be handled in the complex structure-based and template-based approaches? Is there any commonality between these two approaches in handling various types of speech variability while representing long-range temporal dependencies and fine phonetic detail in the overall speech pattern? These and related questions will be discussed by the authors of the invited and contributed papers to be presented in this special session. It is our hope that such focused discussions will enable us to acquire new insights in how to overcome the limitations of HMMs, which in the long run will lead to better functioning ASR systems.

5. References

The references below are listed in alphabetical order.

[1] G. Aradilla, J. Vepa, and H. Bourlard (2005) "Improving speech recognition using a data-driven approach," Proc. EUROSPEECH, Lisbon, pp. 3333-3336.

[2] S. Axelrod and B. Maison (2004) "Combination of hidden Markov models with dynamic time warping for speech recognition," Proc. IEEE ICASSP, Vol. 1, pp. 173-176.

[3] J. Bridle, L. Deng, J. Picone et al. (1998) "An investigation of segmental hidden dynamic models of speech coarticulation for automatic speech recognition," Final Report for the 1998 Workshop on Language Engineering, Johns Hopkins University, 1998, pp. 1-61.

[4] H. Bourlard, H. Hermansky, N. Morgan (1996) "Towards Increasing Speech Recognition Error Rates". Speech Communication, vol. 18, no. 3: 205-231.

[5] L. Deng, M. Aksmanovic, D. Sun, and J. Wu (1994) "Speech recognition using hidden Markov models with polynomial regression functions as nonstationary states," IEEE Trans. Speech & Audio Proc., Vol. 2, pp. 507-520.

[6] L. Deng, D. Yu, and A. Acero (2006) "Structured speech modeling," IEEE Transactions on Audio, Speech and Language Processing (Special Issue on Rich Transcription), Vol. 14, No. 5, pp. 1492-1504.

[7] L. Deng (2006) DYNAMIC SPEECH MODELS --- Theory, Algorithms, and Applications, Morgan & Claypool Publishers.

[8] L. Deng and D. O'Shaughnessy (2003) SPEECH PROCESSING --- A Dynamic and Optimization-Oriented Approach, Publisher: Marcel Dekker Inc., New York, NY.

[9] L. Deng and J. Ma (2000) "Spontaneous Speech Recognition Using a Statistical Coarticulatory Model for the Vocal Tract Resonance Dynamics," J. Acoust. Soc. America, Vol.108, No. 6, Dec 2000, pp. 3036-3048.

[10] M. De Wachter, K. Demuynck, P. Wambacq and D. Van Compernolle (2004) "A Locally Weighted Distance Measure for Example Based Speech Recognition. In Proc. ICASSP, pages 181--184, Montreal, Canada.

[11] M. De Wachter, K. Demuynck, D. Van Compernolle (2006) "Boosting HMM performance with a memory upgrade," Proc. INTERSPEECH, Pittsburgh, pp. 1730-1733.

[12] J. Frankel and S. King (2007) "Speech recognition using linear dynamic models," IEEE Transactions on Audio, Speech and Language Processing, Vol. 15, No.1, pp. 246-256.

[13] O. Ghitza, M.M. Sondhi (1993) "Hidden Markov models with templates as nonstationary states: An application to

speech recognition." Computer, Speech and Language, 7:101-119.

[14] S.D. Goldinger (1996) "Words and voices: episodic traces in spoken word identification and recognition memory." J. of Experimental Psychology; Learning Memory and Cognition, 33: 1166-1183.

[15] S.D. Goldinger (1997) "Words and voices: perception and production in an episodic lexicon." In: K. Johnson & J.W. Mullenix (Eds.), Talker Variability in Speech Processing. Academic Press: 33-66.

[16] Y. Han and L. Boves (2006) "Syllable-length path mixture hidden Markov models with trajectory clustering for continuous speech recognition," Proc. INTERSPEECH, Pittsburgh, pp. 1718-1721.

[17] S. Hawkins (2003) "Contribution of fine phonetic detail to speech understanding." Proc. of the 15th Int. Congress of Phonetic Sciences (ICPhS-03), Barcelona, Spain: 293-296.

[18] S. Hawkins, R. Smith (2001) "Polysp: A polysystemic, phonetically-rich approach to speech understanding. Italian Journal of Linguistics—Rivista di Linguistica," vol. 13: 99-188.

[19] D.L. Hintzman, R. Block, N. Inskeep (1972) "Memory for mode of input." J. of Verbal Learning and Verbal Behavior, 11: 741-749.

[20] L. Jacoby and C. Hayman (1987) "Specific visual transfer in word identification." J. of Experimental Psychology: Learning, Memory,and Cognition, 13: 456-463.

[21] K. Johnson (1991) "Differential effects of speaker and vowel variability on fricative perception." Language and Speech, 34: 265-279.

[22] K. Johnson, J. Mullennix (Eds.) (1997) Talker Variability in Speech Processing. San Diego: Academic Press.

[23] C.-H. Lee (2003) "On Automatic Speech Recognition at the Dawn of the 21st Century," IEICE Trans. on Information and Systems, Special Issue on Speech Information Processing, Vol.E86-D, No. 3, pp. 377-396.

[24] P.A. Luce, C.T. McLennan (2003) "Spoken Word Recognition: The Challenge of Variation." In: C. T. McLennan, P. A. Luce, G. Mauner, & J. Charles-Luce (Eds.), University at Buffalo Working Papers on Language and Perception, 2: 203-240.

[25] V. Maier, R.K. Moore (2005) "An investigation into a simulation of episodic memory for automatic speech recognition." Proc. of Interspeech-2005, Lisbon, 5-9, pp. 1245-1248.

[26] M. Matton, M. Wachter, D. Van Compernolle, and R. Cools (2004) "A discriminative locally weighted distance measure for speaker independent template based speech recognition," Proc. ICSLP, pp. 429-432.

[27] M. Matton, M. De Wachter, D. Van Compernolle and R. Cools (2005) "Maximum Mutual Information Training of Distance Measures for Template Based Speech Recognition". In Proc. International Conference on Speech and Computer, Patras, Greece, October 2005, pp. 511-514.

[28] J.M. McQueen, A. Cutler (2001) "Spoken word access processes: An introduction." Language and Cognitive Processes, 16(5/6): 469-490.

[29] J.M. McQueen, A. Cutler, D. Norris (2003) "Flow of information in the spoken word recognition system." Speech Communication 41: 257-270.

[30] R.K. Moore, A. Cutler (2001) "Constraints on theories of human vs. machine recognition of speech." In: R. Smits, J. Kingston, T.M. Nearey, & R. Zondervan (Eds.), Proc. of SPRAAC (Workshop on Speech Recognition as Pattern Classification), MPI, Nijmegen: 145-150.

[31] M. Ostendorf (1999) "Moving beyond the 'beads-on-a-string' model of speech." Proc. of the IEEE Automatic Speech Recognition and Understanding (ASRU) Workshop, Keystone, Colorado, USA.

[32] M. Ostendorf, V. Digalakis, and J. Rohlicek (1996) "From HMMs to segment models: A unified view of stochastic modeling for speech recognition", IEEE Trans. Speech Audio Proc., Vol. 4, pp. 360-378.

[33] D.B. Pisoni (1997) "Some thoughts on "Normalization" in speech perception." In: K. Johnson & J.W. Mullennix (Eds.), Talker Variability in Speech Processing. Academic Press: 9-32.

[34] L. C. W. Pols (1999) "Flexible, robust, and efficient human speech processing versus present-day speech technology." Proc. of the 14th Int. Congress of Phonetic Sciences (ICPhS-99), San Francisco, USA: 9-16.

[35] A.P. Salverda, D. Dahan, J.M. Mcqueen (2003) "The role of prosodic boundaries in the resolution of lexical embedding in speech comprehension." Cognition, 90: 51-89.

[36] H. Strik (2001) "Pronunciation adaptation at the lexical level." In: J-C. Juncqua, C. Wellekens (Eds.) Proc. of the ITRW 'Adaptation Methods For Speech Recognition', Sophia-Antipolis, France: 123-131.

[37] H. Strik. "Speech is like a box of chocolates." Proceedings of 15th ICPhS, Barcelona, Spain, 2003, pp. 227-230.

[38] H. Strik (2006) "How to handle pronunciation variation in ASR: by storing episodes in memory?" Proc. ITRW on Speech Recognition and Intrinsic Variation (SRIV2006), Toulouse, France, pp. 33-38.

[39] H. Strik, C. Cucchiarini (1999) "Modeling pronunciation variation for ASR: a survey of the literature". Speech Communication, Vol. 29 (2-4), pp. 225-246.

[40] R. Togneri and L. Deng (2003) "Joint State and Parameter Estimation for a Target-Directed Nonlinear Dynamic System Model," IEEE Trans. on Signal Processing. Vol. 51, No. 12, pp. 3061-3070.

[41] D. Yu, L. Deng, and A. Acero (2006) "A lattice search technique for long-contextual-span hidden trajectory model of speech," Speech Communication, Vol. 48, No. 9, pp. 1214-1226.

[42] D. Yu, L. Deng, and A. Acero (2007) "Speaker-adaptive learning of resonance targets in a hidden trajectory model of speech coarticulation," Computer Speech and Language. Vol. 27, pp. 72-87.