Attribute-based Mandarin Speech Recognition using Conditional Random Fields

Chi-Yueh Lin, Hsiao-Chuan Wang

Department of Electrical Engineering, National Tsing-Hua University, Hsinchu, Taiwan

d913920@oz.nthu.edu.tw, hcwang@ee.nthu.edu.tw

10.21437/Interspeech.2007-511

Abstract

Integrating phonetic knowledge into a speech recognizer is a possible way to further improve the performance of conventional HMM-based speech recognition methods. This paper presents a cascaded architecture which consists of attribute detection and conditional random field to make use of phonetic knowledge within the phone decoding process. The attribute detection can be implemented by using any effective feature extraction approaches. In this study, an HMM-based method is applied for attribute tagging of Mandarin speech. Then a conditional random field method which applies attribute labels as the input vectors is used to perform the speech recognition. The preliminary experiment result shows that the proposed method is very promising and worthy for further investigation.

Index Terms: speech recognition, conditional random field, speech attributes

1. Introduction

Conventional speech recognition system using hidden Markov models (HMM) has been the most popular technique for many years. Solid mathematical background and maturity of training/decoding algorithms behind HMM make the speech recognition practical in some applications. Even under a noisy condition, many novel methods have been proposed to make HMM still provide a satisfactory performance. As being noticed in [1], given enough hidden states and a sufficiently rich class of observations, a HMM can accurately model any real-world probability distributions. However, accuracy of HMM heavily relies on the assumption of conditional independence. Roughly speaking, observation X at time t is independent of all previous observations and all previous hidden states for a given state variable Q at time t. This statement is quite strong so that for speech signal in real world the condition is seldom met. This may be one of the reasons that the recognition rate of a HMM-based system improved slowly in recent years. How to break this bottleneck seems to be a challenging and worth going research.

In recent researches, the utilization of phonetic knowledge, such as speech attributes, distinct features, et al., into speech recognition has been highlighted [2][3]. A target phone is not recognized directly from a HMM, instead using several speech evidences/events surrounding the current time to determine which phone is present. Under this procedure, the first process is made of many speech event detectors running in parallel or sequentially. Then a backend decision center gathers all the outcomes produced by these speech event detectors and applies phonetic knowledge on them to finish the decoding process. This kind of procedure has excited many open issues. Take front-end event detectors for example, there exist many techniques and architectures for extracting the speech attributes. Also, we may use different detectors for different events instead of an identical detector for all events. For example of the burst and transient in a stop sound, the conventional framing configuration is not suitable to detect such a short speech event. It is believed that samplebased techniques are much suitable for this case. Given the results provided by the preceding event detectors, how to incorporate these observed events to give the final recognition result is another interesting issue. This problem is similar to the fusion process in other research domains.

In this paper, two modeling techniques are cascaded to achieve a complete decoding process from speech signal to phone labels. The transformation of speech signal into intermediate attributes/events labels is done by conventional HMMs. In other words, all the speech attributes are detected by HMM technique. Then a recent proposed approach, called conditional random fields (CRF) [4], is utilized to incorporate those attributes labels to infer the phone presence. The detail of this cascaded system is given in the next section.

2. HMM/CRF Recognition System

The Mandarin speech recognizer proposed here is composed of two main parts. The front-end HMM-based subsystem consists of a parallel bank of attribute-HMMs. Its function is to transform the input speech frame into several attribute labels. Then the following backend CRF subsystem makes use of these attribute labels to determine the final recognition results. Block diagram of this cascaded system is shown in Figure 1.

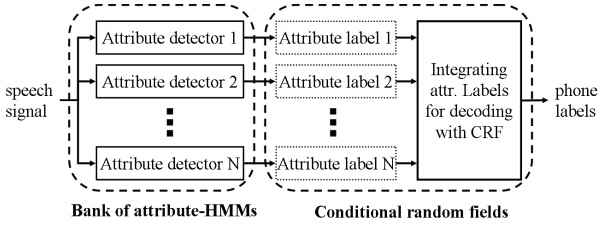

Figure 1: *HMM/CRF system.*

The proposed speech recognition system consists of a cascaded HMM/CRF architecture. The front-end attribute detectors are mainly based on HMMs, then those attribute labels are integrated by the following conditional random fields to give the final recognition results. The detail descriptions of two subsystems are as follows.

2.1. Front-end Attribute-HMMs

In conventional Mandarin speech recognition, the basic units chosen to be modeled are initials and finals. A typical Mandarin syllable consists of one initial and followed by a final. If taking context dependence into consideration, one initial will have many context-dependent variants depending on the type of its following final. However, all finals are still set to be context independent. This special kind of contextdependent models is called right-context dependent (RCD) model.

Here we design a set of attribute-HMMs which are initialized by existing RCD HMMs. For faster training, those RCD HMMs belonging to the same attribute are tied together and re-estimated several times using training data. In our RCD models, each initial is modeled as a left-to-right 3-state HMM; whereas each final is modeled as a left-to-right 5-state HMM. In some circumstance, some initials and finals are both pooled together for model tying, such as the model of N/A (Not Available). Since the numbers of states used in these two kinds of models are different, in practice those initials and finals belonging to the same attribute are pooled and tying separately. In other words, we intrinsically have two N/A models: one for initials and one for finals. During the recognition phase, the output will be the same, however.

We manually design six speech attributes for each Mandarin initial and final. Speech attributes chosen here are manner, final onset type, final ending type, place, aspiration, and voiced. One final can be determined uniquely by three speech attributes: manner, final onset type, and final ending type. On the other hand, one initial can be determined by manner, place, aspiration, and voiced. Table 1 shows HMMs for each attribute.

Table 1: *HMMs in each attribute.*

Attribute	HMMs
manner (Final)	Finals ended with static vocal tract, Finals ended with a vowel, Finals ended with a nasal
manner (Initial)	affricate, fricative, stop, nasal, lateral
Final onset type	/a/, /e/, /o/, /yi/, /yu/, /wu/, /eh/, N/A.
Final ending type	/ai/, /ei/, /yi/, /ao/, /ou/, /wu/, /an/, /en/, /ang/, /eng/, /yu/, /a/, /o/, /e/, /eh/, N/A
place	bilabial, labial-dental, front coronal, middle coronal, back coronal, dorsum, back (velar), N/A.
aspiration	aspiration, non-aspiration, N/A
voiced	voiced, unvoiced

2.1.1. Manner

In this set, we define 3 HMMs for finals and 5 HMMs for initials. Three HMMs for finals are finals ended with static vocal tract configuration, finals ended with a vowel, and finals ended with a nasal. This kind of classification is in a broad phonetic point of view, as comparing to the final ending types which we will introduce later. Five HMMs for initials are affricate, fricative, stop, nasal, and lateral.

2.1.2. Final onset types

This attribute classifies all finals into several groups according to its onset pronunciation type. The onset portion of Mandarin finals should be pronounced with /a/ (ㄚ), /e/ (ㄜ), /o/ (ㄛ), /yi/ (ㄧ), /yu/ (ㄩ), /wu/ (ㄨ), or /eh/ (ㄝ). Sounds not belonging to finals are classified into a special cluster called N/A.

2.1.3. Final ending types

Similar to final onset types, the ending portion of a Mandarin final should be pronounced with /ai/ (ㄞ), /ei/ (ㄟ), /yi/ (ㄧ), /ao/ (ㄠ), /ou/ (ㄡ), /wu/ (ㄨ), /an/ (ㄢ), /en/ (ㄣ), /ang/ (ㄤ),/eng/ (ㄥ), /yu/ (ㄩ), /a/ (ㄚ), /o/ (ㄛ), /e/ (ㄜ), or /eh/ (ㄝ).Compare to final onset types, ending types are more complicated since Mandarin finals including simple finals and compound finals. Articulators are fixed while pronouncing in the former case, however, in the latter case the articulator configurations are changing with time. As mentioned above, there is still a N/A cluster for those sounds not belonging to finals.

2.1.4. Place

Each Mandarin initial can be classified into one of the following groups: bilabial, labial-dental, front coronal, middle coronal, back coronal, dorsum, or back (velar). Sounds belonging to finals are grouped as N/A.

2.1.5. Aspiration

Aspiration and Non-Aspiration attributes are mainly for stops and affricates, all the other initials and all the finals are grouped into N/A.

2.1.6. Voiced

This is the most simple and straightforward attribute. All finals are voiced, whereas all initials but /m/ (ㄇ), /n/ (ㄋ), /l/ (ㄌ), and /r/ (ㄖ) are unvoiced.

Each input utterance is framed with 25ms frame size and in 10ms frame rate. Then each frame is parameterized to a 42- dimentional feature vector that including 13-order MFCC, normalized log-energy, and their corresponding delta and delta-delta coefficients. Techniques of pre-emphasis and cepstral mean subtraction are also applied in the pre-processing. Six sets of attributes-HMMs described above run in parallel to decode the utterance into six attribute representations. In other words, the output of each frame after this decoding phase will be six attributes labels. These attribute labels are treated as intermediate representations and are then processed by the subsequent subsystem which using conditional random fields to transform the attribute labels into their final representation form, Mandarin initials or finals.

2.2. Backend Conditional Random Fields

Conditional Random Fields (CRFs) [4] are undirected graphical models. The main advantage of CRFs is their flexibility to include a wide variety of arbitrary, nonindependent features. It has been widely applied to the field of natural language processing, especially in the task of POS tagging and NP-chunking. According to the Hammersley- Clifford theorem, CRFs define the conditional probability of a set of output values given a set of input values to be proportional to the product of potential functions on cliques of the graph,

$$P_\Lambda(\mathbf{s}|\mathbf{o}) = \frac{1}{Z_o} \prod_{c \in C(\mathbf{s},\mathbf{o})} \Phi_c(\mathbf{s}_c, \mathbf{o}_c) \qquad (1)$$

where $\Phi_c(\mathbf{s}_c, \mathbf{o}_c)$ is the clique potential on clique c, and Z_o is a normalization factor over all output values. Z_o is also known

as partition function.

Among the family of CRFs, the most popular one is linear-chain CRF, which can be expressed as following,

$$p(\mathbf{s}|\mathbf{o}) = \frac{1}{Z(\mathbf{o})} \exp\left(\sum_{t=1}^{T}\sum_{k=1}^{K}\lambda_k f_k(s_{t-1}, s_t, \mathbf{o}, t)\right) \quad (2)$$

where $\mathbf{o} = \{o_1, o_2, \cdots, o_T\}$ is the observation vector spanning T frames, $\mathbf{s} = \{s_1, s_2, \cdots, s_T\}$ is its corresponding state sequence, $f_k(s_{t-1}, s_t, \mathbf{o}, t)$ represents the k-th feature function that measures the degree to which state s_t is compatible with a transition from state s_{t-1} and with observation $\mathbf{o}$, λ_k is the weighting factor for feature function f_k, and $Z(\mathbf{o})$ is a normalization constant. In practice, feature function f_k can be designed manually or generated automatically [5] for domain-specific applications. In order to tell the difference between HMM and CRF, remind that in CRF the transition between adjacent states depends on the whole observations and we have no assumption about the distribution over observations.

While training CRF, weighting factor λ_ks which reflecting the importance of feature function f_k are the only parameters to be considered. Given a set of training samples of the form $(\mathbf{o}_i, \mathbf{s}_i)$, where $\mathbf{o}_i = (o_{i,1}, o_{i,2}, \cdots, o_{i,T})$ and $\mathbf{s}_i = (s_{i,1}, s_{i,2}, \cdots, s_{i,T})$, the CRF can be trained to maximize the log-likelihood of the training data, possibly with a regularization penalty to prevent overfitting. Let $\Theta = \{\lambda_1, \cdots, \lambda_K\}$ denote all the tunable parameters in the model. Then the objective function to be maximized is

$$\begin{aligned} J(\Theta) &= \log\prod_{i=1}^{N} p(\mathbf{s}_i|\mathbf{o}_i) - \sum_{k=1}^{K}\frac{{\lambda_k}^2}{2\sigma^2} \quad (3)\\ &= \sum_{i=1}^{N}\sum_{t=1}^{T}\sum_{k=1}^{K} f_k(s_{i,t-1}, s_{i,t}, \mathbf{o}_i, t) - \sum_{i=1}^{N} Z(\mathbf{o}_i) \\ &\quad - \sum_{k=1}^{K}\frac{{\lambda_k}^2}{2\sigma^2} \end{aligned}$$

The last term of the right-hand side of Eq. (3) is a regularization penalty which can be viewed as performing MAP estimation of Θ, if Θ is assigned a Gaussian prior with mean 0 and covariance $\sigma^2\mathbf{I}$. The partial derivative of objective function $J(\Theta)$ is

$$\begin{aligned} \frac{\partial J(\Theta)}{\partial \lambda_k} &= \sum_{i=1}^{N}\sum_{t=1}^{T} f_k(s_{i,t-1}, s_{i,t}, \mathbf{o}_i, t) \quad (4)\\ &\quad - \sum_{i=1}^{N}\sum_{t=1}^{T}\sum_{s,s'} f_k(s, s', \mathbf{o}_i, t)p(s, s'|\mathbf{o}_i) - \frac{\lambda_k}{\sigma^2} \end{aligned}$$

Some gradient ascent methods like iterative scaling and L-BFGS can be applied to find the optimum value of each λ_k. Here we choose L-BFGS as our training algorithm since it is faster than iterative scaling and conjugate gradient [6].

To incorporate CRF in our task, observation vectors $\mathbf{o}$ are attribute labels generated from the attribute-HMMs and state variables $\mathbf{s}$ are recognized initial/final labels. At time t, we determine the output initial/final label s_t according to its nearby observations at time t. How these nearby observations are interacted to determine the current output label is described by feature functions and their corresponding weighting factors. This flexibility of including observations over a longer time span is one advantage of CRF over HMM. The feature functions designed here are described as follows,

- Manner, final onset type, and final ending type at current frame, previous frame, and next frame.
- Manner, place, aspiration, and voiced at current frame, previous frame, and next frame.
- Six attributes individually at current frame, previous frame, and next frame.

In other words, we apply a sliding window of length three centered at each time step over observation $\mathbf{o}$. Transition from previous state s_{t-1} to current state s_t actually depends on observations o_{t-1}, o_t, o_{t+1}, not the entire observation. Apply a wider sliding window would include more observation vectors into consideration, though complexity will increase during training and inference. For simplicity, feature functions used here are chosen to be binary-valued, though they can be real-valued. The training and decoding process of conditional random fields are with help of CRF++ package [7].

3. Experiments

Six sets of attribute-HMMs are trained on TCC300 Mandarin speech corpus which was collected by three universities in Taiwan. Each utterance in the corpus was recorded in 16-bit PCM format with 16kHz sampling rate under office environments. The collected utterances were divided into two sets, training set and evaluation set. In the training set, there are 24,742 utterances and the total length is about 24 hours. On the other hand, there are 2,595 utterances in the evaluation set and roughly 2.5 hours in total. Since the computation of training attribute-HMMs directly from the raw speech data is very expensive, the attribute-HMMs are initialized from RCD-HMMs as described in section 2. Table 2 gives the recognition results of each set of attribute-HMMs.

Table 2: *Recognition results of six sets of attribute- HMMs.*

Attributes	Correction(%)	Accuracy(%)
manner	81.39	76.34
Final onset type	86.12	82.59
Final ending type	87.14	82.89
place	83.82	80.77
aspiration	88.61	84.72
voiced	86.71	83.93

With further analysis of each performance in attribute-HMMs, we find some unsatisfactory results. In attribute manner, the most misrecognized one is fricative (about 71% in correction rate). Most of time fricatives are misclassified into affricates or stops. In attribute place, coronals are often misclassified into each other, making some initials hard to be classified. For final ending type, finals ending with a nasal sound would be easily confused, which is corresponding to the most common happened pronunciation errors while speaking Mandarin. Misclassification of the front-end attributes would influence the decoding performance of the backend CRF since our feature functions are chosen to be binary-valued. Mistakes in the attribute labels would reduce the inference accuracy of CRF.

Table 3 gives the simple comparison of conventional HMM based speech recognition method and the proposed backend CRF decoding process. In CI-HMMs, there exist 22 context-independent initial models and 39 context-independent models. However, in RCD-HMMs case, initial models are extended to 99 according to its right context. The number of final models in RCD-HMMs does not change since the context independency is still retained. Notice that the results given in the table are decoding with free grammar. No constraints are made in the decoding process. We expect that the results will be better if we add linguistic constraints in the decoding phase. The most promising result comes from the accuracy rates given by CRF, though the correction rate is lower than that of CI-HMMs and RCD-HMMs. The low accuracy rate in RCD-HMMs case is due to its higher insertion rate caused by more trained models compared to other systems. We also find that the selection of feature functions in CRF will influence the decoding results dramatically. The more distinct features included in feature functions, the more reliable results are given.

Table 3: *Results of Mandarin initials/finals recognition task by three different systems. No linguistic constraints are applied during the decoding phase.*

System	Correction(%)	Accuracy(%)
CI-HMM	69.61	54.91
RCD-HMM	71.84	44.12
HMM/CRF	61.60	58.25

4. Discussion

In this paper, we use hidden Markov model as our attribute detector. HMM is a kind of generative models which is a counter part of discriminative models. Some have argued that discriminative models would be better than generative models in classification tasks. However, HMMs are preferred here because its capability of including the influence of hidden variables and manipulating the data in a segmental way. Moreover, solid mathematical background behind HMM is another reason makes it so attractive and appealing. As mentioned earlier, how to design a specific speech attribute detector is an open issue. Detectors from the same family are almost impossible to model or classify all kinds of speech attributes well equally. To pursue a higher detection rate, HMM may not the only choice. One should select and design a specific detector according to the uniqueness of the target speech attribute. The more reliable the front-end, the better recognition result can be given by the following backend system.

In the backend system, the preceding attribute labels are incorporated to infer the present phone by a linear-chain CRF. CRFs make inference by incorporating attribute evidences in an exponential framework and weight the importance of individual evidence by its corresponding weighting factor. CRFs have been successfully applied to many applications, like POS tagging and NP-chunking tasks in natural language processing, or in the tasks of image and video segmentation. Among CRF family, linear-chain CRF is the most concise one. There exist other types of CRF, like semi-Markov CRF [8], hidden-CRF, and Dynamic-CRF [9]. There could be other candidates for speech processing. Especially, in a graphical model point of view, the graphical topology of hidden-CRF is almost identical to HMM except the hidden-CRF is an undirected model. It reserves the characteristics of including hidden variables and treating the states and observations in a segmental way. In the task of phone classification [10] and gesture recognition [11], hidden-CRF has shown its promising applicability.

5. Future Work

In this paper, we have implemented a prototype of attribute-based speech recognition system. There still exist many open issues to be addressed under this framework. The front-end detectors should be further investigated to find out which kind of detector is more suitable for some specific speech attribute, and how many attributes should be included to identify a target phone. For the backend system, linear-chain CRF can be replaced by other types of CRF to more accord with speech characteristics. In such way, the attribute-based speech recognition system is a very promising approach to break the bottleneck faced nowadays.

6. Acknowledgements

This research was sponsored by the National Science Council, Taiwan, under contract number NSC-95-2221-E-007-217.

7. References

[1] A. Bilmes, "What HMMs can do", IEICE Trans. on Information and Systems, Vol. E89-D, No. 3, March 2006.

[2] J. Morris, and E. Fosler-Lussier, "Combining Phonetic Attributes using Conditional Random Fields," in Proceedings of INTERSPEECH 2006, pp. 597-600.

[3] H. Yu, and A. Waibel, "Integrating Thumbnail Features for Speech Recognition using Conditional Exponential Models," in Proceedings of ICASSP 2004, pp. 893-896.

[4] J. Lafferty, A. McCallum, and F. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," in Proceedings of International Conference of Machine Learning, 2001.

[5] A. McCallum, "Efficiently Inducing Features of Conditional Random Fields," in Proceedings of Uncertainty in Artificial Intelligence, 2003.

[6] F. Sha, and F. Pereira, "Shallow Parsing with Conditional Random Fields," in Proceedings of Human Language Technology, NAACL, 2003.

[7] T. Kudo, "CRF++: Yet Another CRF toolkit," in http://chasen.org/ taku/software/CRF++/

[8] S. Sarawagi, and W.W. Cohen, "Semi-Markov Conditional Random Fields for Information Extraction," in Proceedings of Advanced in Neural Information Processing Systems (NIPS), 2004.

[9] C. Sutton, K. Rohanimanesh, and A. McCallum, "Dynamic Conditional Random Fields: Factorized Probabilistic Models for Labeling and Segmenting Sequence Data," in Proceedings of International Conference on Machine Learning, 2004.

[10] A. Gunawardana, M. Mahajan, A. Acero, and J.C. Platt, "Hidden Conditional Random Fields for Phone Classification," in Proceeding of INTERSPEECH 2005, pp. 1117-1120.

[11] L. Morency, "Context-based Visual Feedback Recognition," Technical report, CSAIL Lab., MIT, 2006.