GENDER RECOGNITION AND GENDER-BASED ACOUSTIC MODEL ADAPTATION FOR TELEPHONE-BASED SPOKEN DIALOG SYSTEM

Kinfe Tadesse Mengistu, Martin Schafföner, Andreas Wendemuth

Cognitive Systems Group, Otto-von-Guericke University
Kinfe.Tadesse@E-Technik.Uni-Magdeburg.de

Abstract: In this paper we describe the speech recognition component of a telephone-based spoken dialog system that uses HTK-based speech recognizer integrated in a VoiceXML framework and an ISDN telephone interface. As the speech recognizer component is one of the most decisive components that determine the usefulness and user acceptance of a dialog system, we present here strategies on how to build and improve the performance of a speech recognition component within such a system. The baseline speaker-independent system gives a word error rate (WER) of 13.66% for female speakers and 21.55% for male speakers using a 22-hour telephone speech from the Communicator 2001 Evaluation corpus. As can be observed, the system appears biased towards female speakers. This is attributed to the fact that the number of female speakers used in training the models is significantly higher than male speakers (72 vs. 28). To combat this problem and to improve the performance of the system for male speakers, we use two approaches. First, taking the presence of within-gender acoustic similarity due to similar vocal mechanism of speakers into consideration, we adapt the speaker independent HMMs using adaptation data from each gender. As an alternative, separate gender-dependent models are built. We also built a Gaussian Mixture Model (GMM) gender classifier that can determine the gender of the speaker given a very short utterance (typically a "yes" or a "no") with 96.62% accuracy.

1 Introduction

A telephone-based spoken dialog system is comprised of a telephone network interface to deliver calls into the system, a speech recognizer to accept requests from users, a text to speech synthesizer (TTS engine) for playing prompts and responses to the caller, a semantic interpreter for comprehending requests, a mechanism for response generation, and a dialog manager to orchestrate the various components.

The speech recognizer in our dialog system uses HTK [1] to build recognition resources and its API (ATK) to build a real-time speech recognizer [2] integrated in a VoiceXML framework. Among other features, ATK allows a flexible use of resources during the recognition process. It uses a global configuration file where HTK compatible HMM models and other recognition resources such as grammar, HMM list, and pronunciation lexicon are specified [2]. This makes it possible to use the same framework for various application domains and languages by simply building the necessary recognition resources offline and specifying them in the configuration file. The choice of an open VoiceXML platform is an important design decision. We have chosen OptimTalk[1] as it is open enough to allow the integration of our own speech recognizer, telephone interface, etc.

[1] http://www.optimsys.cz/

As HTK-based speech recognizers require grammar in HTK's Standard Lattice Format, a separate grammar component supporting this grammar format within the VoiceXML framework is developed. Our ATK-based semantic interpreter simply ignores semantically irrelevant terms from the recognition output and parses only the content bearing terms. The use of grammars specific to a given dialog state and ignoring irrelevant filler-words improves the performance of the system further in real time. The system combines the power and flexibility of HMM-based speech recognizer and the convenience of VoiceXML for dialog authoring.

2 Data Preparation

The data used in this endeavor consists of a total of 22 hours of telephone speech from the Communicator 2001 Evaluation corpus [3]. The corpus consists of utterances recorded as users interacted with eight[2] different Airline Travel Planning dialog systems. We split the 22-hour telephone speech data spoken by 149 speakers into five sets; namely, the training set which consists of 15 hours of speech (12,863 utterances) spoken by 100 speakers (28 male and 72 female) to build the baseline speaker independent model; male adaptation set (526 utterances spoken by 6 male speakers) to adapt the speaker independent model to male speech; male test set (1047 utterances spoken by 8 male speakers) to test how well the various models perform for male speakers; female adaptation set (2522 utterances spoken by 15 female speakers), and female test set (2892 utterances spoken by 20 female speakers) to test the performance of the various models for female speakers. We use a merger of the male and female adaptation data as a development set in order to determine certain parameters experimentally.

For gender identification, the same acoustic training data is used to train the gender recognizer, but for testing purposes we merged the adaptation and testing data of both male and female speakers and selected only 1450 short utterances which are mainly “yes”, “yeah”, and “no” spoken by all speakers (14 male and 35 female speakers) as we would like to decide the gender of the speaker with the first utterance which is essentially a yes or a no in our dialog design.

As the vocabulary used in the application domain is fairly limited (about 1200 distinct words), a back-off bigram language model is built on the training transcriptions. The pronunciation lexicon is based on the CMU public domain pronunciation dictionary.

3 Feature Extraction

We used Mel-Frequency Cepstral Coefficients (MFCCs), which are widely used features for automatic speech recognition systems [4] to transform the speech waveform into a sequence of discrete acoustic vectors. The MFCCs are computed by performing pre-emphasis on the acoustic waveform, dividing the incoming waveform into blocks of 25ms length and 10ms overlap, multiplying each block by a Hamming Window, followed by removing the DC offset from each windowed excerpt of the waveform. Then the Fast Fourier Transform (FFT) of the windowed signal is calculated and the square of the magnitude (i.e., the power spectrum) is fed into a series of Mel-Frequency filterbank channels. Then, Discrete Cosine Transform (DCT) is applied to the logarithm of the filterbank outputs. The Discrete Cosine Transform has a notable effect in favor of the diagonal covariance assumption by de-correlating the features in the feature vectors so that each feature can be assumed to be independent of any other feature. Finally, the first and second

2 ATT, BBN, Carnegie Mellon University, IBM, Lucent Bell Labs, MIT, SRI and University of Colorado at Boulder

order temporal time differences (i.e. the differences between parameter values over successive frames - delta, and delta-delta coefficients) are computed to better model temporal variation of the speech spectrum. The overall output of this process is a feature vector containing 39 components made up of 13 cepstral coefficients including the 0^{th} order coefficient (c_0 to c_{12}) and the corresponding delta and delta-delta coefficients.

4 The Baseline System

4.1 Acoustic Modeling

The purpose of acoustic modeling is to estimate the state transition probabilities and the observation likelihood of an observation vector given an HMM state. We use the Baum-Welch algorithm that is an implementation of Expectation Maximization in order to estimate these parameters given the training data in the form of observation vectors.

A weighted mixture of multivariate Gaussians is used to model observation likelihoods. The acoustic likelihood $b_j(o_t)$ of a feature vector o_t given an HMM state j with mean vectors μ_{jm} and diagonal covariance values σ^2_{jmd} is given by:

$$b_j(o_t) = \sum_{m=1}^{M} \frac{\omega_{jm}}{\sqrt{(2\pi)^D \prod_{d=1}^{D} \sigma^2_{jmd}}} \exp\left(-\frac{1}{2}\sum_{d=1}^{D} \frac{(o_{dt} - \mu_{jmd})^2}{\sigma^2_{jmd}}\right)$$

where D is the dimension of the feature vectors, M is the number of Gaussian components per state, and ω_{jm} is the mixture weight of the m^{th} component in state j satisfying the property $\sum_{m=1}^{M} \omega_{jm} = 1$.

4.2 Basic Context Independent Models

We represent each monophone by a hidden Markov model of 3 emitting states with left-to-right topology, where each emitting state has two transitions: back to itself and to the next state. The entry and exit states are non-emitting and consequently have no probability output distribution associated with them. The mean and variance of each state of the monophone models are initialized to the global mean and variance of the training data. The transition probabilities from an emitting state back to itself and to the next state are set equiprobable; the transition from the entry state to the first emitting state is set to 1.0 and all other transitions are set to zero.

The parameters of the models are then re-estimated in two consecutive runs of the Baum-Welch algorithm using monophone transcription of the training data. Then the silence model is made to take care of impulsive noises in the training data by adding extra transitions from state 2 to 4 and from state 4 to 2 in the model. Furthermore, to account for any pauses introduced by the speaker between words of an utterance, a one state short pause (sp) model is created whose emitting state is tied to the center state of the silence model. Then two more iterations of the Baum-Welch algorithm are run on the resulting models. To account for multiple pronunciations of some words in the dictionary, a new, realigned transcription is generated that contains the pronunciation that best matches the acoustic evidence using the Viterbi alignment.

4.3 Context Dependent Models

Context-independent models are poor discriminators as a phoneme can be realized differently depending on context. The common approach to achieve good phonetic discrimination is to use triphones. Crossword triphones (XWRD) where context spans word boundaries are considered.

In order to build context-dependent, tied-state triphone models, we start with a set of basic context-independent monophone HMMs trained as described in section 4.2. The resulting single-Gaussian monophone models are then used to generate triphone prototypes and re-estimated using the Baum-Welch algorithm with a triphone list and triphone transcriptions.

When triphones are used, usually training data becomes insufficient as the resulting system has too many models to train. Tying is one way to deal with this problem of data insufficiency. We use a phonetic decision tree that is based on asking phonetic questions about the left and right contexts of each triphone to distinguish clusters and tie similar states within triphone sets.

The stop criteria for clustering are the outlier threshold that determines the minimum number of training data that each leaf in the decision tree must have to stand as a cluster, and the threshold specifying the increase in log likelihood that has to be achieved by any question at any node. If a split in the decision tree increases the log-likelihood by less than this value, splitting stops and the decision tree is complete. The optimal values for these two parameters are experimentally determined using the development set.

Once we have single-Gaussian, tied-state word-internal or crossword triphones, we increment the number of Gaussian mixture components to the desired number. It was experimentally found that 32 Gaussian mixture components per state are optimal for this setup using the development set. The number of Gaussian components per state is incremented by cloning the component with the largest mixture weight, dividing the component weight by 2, and perturbing the means by +/- 0.2σ. The resulting models are then re-estimated with 8 consecutive runs of the Baum-Welch algorithm. This is repeated until we have estimated models with the required number of mixtures.

5 Gender Recognition

In an utterance, not only the message that the speaker wants to express but also hidden information that include the speaker's age group, gender, and other speaker dependent information are conveyed. Considering gender, it can be observed that there is apparent difference between the mean and variance of male and female feature vectors. Therefore, it is possible to use a Gaussian Mixture Model (GMM) to identify the gender of a speaker given the parameters of an utterance spoken by a male or female speaker. In order to use the existing HTK infrastructure for gender recognition, we modeled a GMM as a single-state hidden Markov model (HMM) with a Gaussian mixture observation density where there is no state transition probability within the model.

In a GMM-based gender classifier, the parameters of an utterance are modeled with the mixture weights, mean vectors and variance parameters of the component densities. The feature vectors are assumed to be independent. Therefore, the log-likelihood of a model λ for a sequence of feature vectors $O = \{o_1, o_2, \ldots, o_T\}$ is defined as:

$$\log p(O \mid \lambda) = \sum_{t=1}^{T} \log p(o_t \mid \lambda)$$

where and $p(o_t \mid \lambda)$ is computed for a D dimensional feature vector according to:

$$p(o_t \mid \lambda) = \sum_{m=1}^{M} \frac{\omega_m}{\sqrt{(2\pi)^D \prod_{d=1}^{D} \sigma_{md}^2}} \exp\left(-\frac{1}{2} \sum_{d=1}^{D} \frac{(o_{dt} - \mu_{md})^2}{\sigma_{md}^2}\right)$$

where M is the number of Gaussian components per model, and ω_m is the mixture weight of the m^{th} component satisfying the property $\sum_{m=1}^{M} \omega_m = 1$.

Two GMMs (one for each gender) are trained using the Baum-Welch algorithm in order to estimate the likelihood of the model λ given a sequence of training data in the form of feature vectors. The experimental results are following in Section 7.1.

6 Gender-Adapted Models

The performance of speaker independent (SI) systems degrades when tested with speakers and environments that are not sufficiently represented in the training corpus. Therefore, it is necessary to adapt SI acoustic models to the new context in which the speech recognizer is to be used. Speaker adaptation uses speaker specific information given in an adaptation data to adjust the acoustic model parameters (mean and variance of the density functions) of the initial speaker independent model to reflect the characteristics of the current speaker. Considering the existence of within-gender acoustic similarity due to similar vocal tract size of speakers of the same gender, adapting the speaker independent HMMs using adaptation data from each gender can give robust gender dependent models. The idea is to capture gender specific characteristics from the adaptation data and transform the model parameters of the initial model set accordingly to get gender-adapted HMMs that could perform better than the original speaker independent model.

The gender-adapted models are built using supervised, Maximum Likelihood Linear Regression (MLLR), and Maximum A Posteriori (MAP) adaptation techniques implemented in HTK.

6.1 Maximum Likelihood Linear Regression (MLLR)

Maximum Likelihood Linear Regression (MLLR) is a transformation-based method that estimates linear transformations for the model parameters (mixture components of HMMs) to maximize the likelihood of the adaptation data [5]. MLLR uses a regression class tree to cluster acoustically similar Gaussians that are close to each other in acoustic space into regression classes that share a common transform. This makes adaptation of densities for which there were no observations in the adaptation data possible [1].

The adaptation of the transition probabilities and the mixture component weight will have little effect on the final performance [6]. However, transformation of the diagonal covariance matrix can give performance improvement. Since reliable variance estimation from a limited amount of data is difficult, only Gaussian mean vectors are updated in the experiments presented here. The experimental results are presented in section 7.4.1.

6.2 Maximum A Posteriori (MAP) Adaptation

Maximum A Posteriori (MAP) estimation is a model-based approach that maximizes the posterior probability using prior knowledge about the model parameter distribution. Given good

initial models and large amount of adaptation data, MAP can perform better than MLLR. MAP re-estimates the models unlike MLLR that transforms the models and re-estimation requires large amount of data. The drawback of MAP approach is that if there was insufficient adaptation data for a phone to reliably estimate a sample mean, no adaptation is performed [7] for that phone. The experimental results are shown in section 7.4.2.

7 Experimental Results

7.1 Gender Recognition

A Gaussian Mixture Model (GMM) based gender recognizer is built using 39 MFCC coefficients extracted as described in section 3. The number of Gaussian mixture components required for adequate performance and the number of iterations between each Gaussian increment were experimentally found to be 32 and 4, respectively on the development set. In order to find out which features and coefficients yield the best result, Perceptual Linear Prediction (PLP) coefficients and MFCC features were tried as shown in Table 1.

As can be seen in Table 1, MFCC feature vectors including C_0 as the energy term give the best result. This was also true for speech recognition in our setup. Traditionally the 0^{th} MFCC coefficient is considered futile and is discarded; however, as described in [8] the 0^{th} coefficient contains a collection of average energies of each frequency band in the signal being analyzed and hence is useful.

MFCC Feature Kind	Accuracy (%)
MFCC_E_D_A	94.0
MFCC_0_D_A	96.62
PLP_0_D_A	96.41
PLP_E_D_A	91.52

Table 1 - GMM-based Gender Classifier

7.2 Speaker Independent Models

To measure the performance of the speaker independent system for male and female users, the baseline system is tested using separate male and female speakers. As can be seen in the following table, the performance of the model for female speakers is by far better than male speakers. This clearly attributed to the fact that male speakers are less represented in the training data.

HMMTYPE	Gender	Accuracy (%)
XWRD	Male	78.45
XWRD	Female	86.34
MONO	Male	72.68
MONO	Female	81.87

Table 2 - The Baseline Speaker Independent Model

7.3 Gender Dependent Models

Building separate gender dependent models for male and female speakers is one way to improve performance as the inter-speaker variability is now limited to a given cluster. As can be seen in Table 3, gender dependent models give apparent performance improvement.

HMMTYPE	Gender	Accuracy (%)
XWRD	Male	78.69
XWRD	Female	86.46
MONO	Male	76.66
MONO	Female	82.72

Table 3 - Gender Dependent Models

7.4 Gender-Adapted Models

We have generated gender-adapted models from the speaker independent model using MLLR, MAP and a combination of the two.

7.4.1 MLLR adaptation of the Means

The most important design parameter to decide is the number of classes required which must be determined empirically. We found 32 and 40 as appropriate number of classes for the male and female models experimentally using the adaptation data of the respective gender. As can be seen in Table 4 the MLLR gender-adapted models give comparable results to the separate gender dependent models shown in Table 3. In both cases, considerable improvement is obtained.

HMMTYPE	Gender	Accuracy (%)
XWRD	Male	79.79
XWRD	Female	86.43
MONO	Male	75.19
MONO	Female	82.51

Table 4 - MLLR Adaptation

7.4.2 MAP adaptation

For MAP adaptation, the model parameter distribution of the speaker independent model is used as the informative priors. An important design parameter is the scaling (relevance) factor (τ), which is a weighting of the prior knowledge to the adaptation speech data. This value has been experimentally found to be 80 and 100 for male and female models, respectively using the adaptation data of the respective gender.

HMMTYPE	Gender	Accuracy (%)
XWRD	Male	78.59
XWRD	Female	86.19
MONO	Male	74.98
MONO	Female	82.25

Table 5 - MAP Adaptation

As can be seen in the Table 5, MAP adaptation also gives improved result for monophone models but not for crossword models. This could be attributed to the fact that the amount of adaptation data is not sufficient to reliably re-estimate the model parameters for crossword triphones.

7.4.2 MLLR and MAP adaptation combined.

As can be seen in Table 6, using the MLLR transformed means as the informative priors for MAP adaptation gives slightly improved results.

HMMTYPE	Gender	Accuracy (%)
XWRD	Male	79.81
XWRD	Female	86.45
MONO	Male	75.33
MONO	Female	82.71

Table 6 *- MLLR and MAP combined*

8 Conclusion

Due to the gender-imbalance in the training corpus, the resulting SI model performed not so well for the under represented gender. We, therefore, described the various adaptation techniques that can be used in order to generate gender-adapted models from speaker independent HMMs. Generating gender dependent models using adaptation technique is preferred to building separate gender dependent models as fully trained initial models contain some general speech information that can be useful for the new system as well. However, the gain in terms of accuracy is not significant, and this may be attributed to the fact that the algorithms, which are essentially speaker adaptation techniques, did not capture enough within-gender similar acoustic information from the adaptation data. A GMM gender recognizer that can decide the gender of a speaker with the first and very short utterance with an accuracy of 96.62% is also reported.

Reference

[1] S. Young, G. Evermann, M. Gales, T. Hain, X. Liu, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valchev, P. Woodland. The HTK Book. Revised for HTK Version 3.4, 2006.

[2] S. Young: ATK. An Application Toolkit for HTK Version 1.6. Cambridge University, http://mi.eng.cam.ac.uk/research/dialogue/ATK_Manual.pdf , 2007.

[3] M. Walker, J. Aberdeen, and G. Sanders: 2001 Communicator Evaluation. Linguistic Data Consortium, Philadelphia, 2003.

[4] S.B. Davis, and P. Mermelstein, P.: Comparison Of Parametric Representation For Monosyllabic Word Recognition In Continuously Spoken Sentences. IEEE Trans. on ASSP, 1980.

[5] M.J.F. Gales and P.C. Woodland: Variance Compensation Within The MLLR Framework, Technical Report 242, Cambridge University Engineering Department, UK, 1996.

[6] C.J. Leggeter and P.C. Woodland: Maximum Likelihood Linear Regression For Speaker Adaptation Of Continuous Density Hidden Markov Models. Computer Speech and Language, vol.9, pp.171–185, 1995.

[7] Z. Wang, T. Schultz, A. Waibel: Comparison of Acoustic Model Adaptation Techniques on Non-native Speech. Proc. ICASSP (2003). pp. 540-543.

[8] F. Zheng, and G. Zhang: Integrating the Energy Information into MFCC, In International Conference on Spoken Language Processing (ICSLP), vol.1, pp. 389-392. 2000