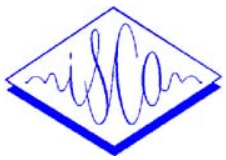

EUROSPEECH '97
5th European Conference on Speech Communication and Technology
Rhodes, Greece, September 22-25, 1997

SOURCE NORMALIZATION TRAINING FOR HMM APPLIED TO NOISY TELEPHONE SPEECH RECOGNITION

Yifan Gong

Speech Research, Media Technologies Laboratory, Texas Instruments
P.O.BOX 655303 MS 8374, Dallas TX 75265, U.S.A.
Email: Yifan.Gong@ti.com

ABSTRACT

We refer to *environment* e as some combination of speaker, handset, transmission channel and noise background condition, and regard any practical situation of a speech recognizer as a mixture of environments.

A speech recognizer may be trained on multi-environment data. It may also need to adapt the trained acoustic models to new conditions. How to train an HMM with multi-environment data and from what seed model to start an adaptation are two questions of great importance.

We propose a new solution to speech recognition which is based on, for both *training* and *adaptation*, a separate modeling of phonetic variation and environment variations. This problem is formulated under hidden Markov process, where we assume,

- Speech x is generated by some canonical (independent of environmental factors) distributions,
- An unknown linear transformation W_e and a bias b_e, specific to environment e, is applied to x with probability P(e),
- x cannot be observed, what we observe is the outcome of the transformation: $o = W_e x + b_e$.

Under maximum-likelihood (ML) criterion, by application of EM algorithm and the extension of Baum's forward and backward variables and algorithm, we obtained a joint solution to the parameters of the canonical distributions, the transformations and the biases, which is novel.

For special cases, on a noisy telephone speech database, the new formulation is compared to per-utterance cepstral mean normalization (CMN) technique and shows more than 20% word error rate improvement.

1. INTRODUCTION

We refer to *environment* as speaker, handset, transmission channel and noise background conditions. Any speech signal can only be obtained in a particular environment. Speech recognizers suffer from environment variability, for two reasons:

- Trained model distributions may be biased from testing signal distributions because of environment mismatch.
- Trained model distributions are flat because they are averaged over different environments.

For the first problem, the environmental mismatch can be reduced through *model adaptation* [6, 5, 7], based on some utterances collected in the testing environment. To solve the second problem, which has not been addressed until recently, the environmental factors should be removed from the speech signal *during the training* procedure, i.e. by *source normalization*.

A practical speech recognizer may be given a variety training data, collected with different speakers, handsets, transmission channels and background noises. It may also adapt the trained acoustic models to new conditions.

Optimality-related questions therefore arise: 1. How to train an HMM with data collected in different conditions? 2. What is the optimum seed model to start an adaptation? Source normalization presented in this paper provides an answer to these questions.

In the direction of source normalization, speaker adaptive training [2] uses linear regression (LR) to decrease inter-speaker variability. Another technique models mean vectors as the sum of a speaker-independent bias and a speaker-dependent vector [1]. However, regarding to environment, both techniques are supervised, i.e.: they require explicit label of the classes, e.g. speaker or gender of the utterance during the training. Therefore they cannot be used to train classes which cannot be labeled, such as acoustically close speakers, handsets or background noises. Such inability of discovering clusters may be a disadvantage in application.

We provide a maximum likelihood (ML) LR solution to the environment normalization problem, where the environment is modeled as a hidden (non-observable) variable. An EM-based training algorithm can generate any number of optimal clusters of environments and therefore it is not necessary to label a database in terms of environment. For special cases, the technique is compared to per-utterance cepstral mean normalization (CMN) technique and shows performance improvement on a noisy telephone speech database.

2. SOURCE NORMALIZATION

2.1. Formulation

We assume: 1. The speech signal $\mathbf{x}$ is generated by continuous density hidden Markov model (CDHMM), called sources. 2. Before being observed, the signal has undergone an environmental transformation, drawn from a set of transformations. Such a transformation is linear, and is

independent of the mixture components of the source. Let $\mathbf{W}_{je}$ be the transformation on the HMM state j of the environment e. 3. There is a bias vector $\mathbf{b}_{ke}$ at the k-th mixture component due to environment e.

What we observe at time t is thus:

$$\mathbf{o}_t = \mathbf{W}_{je}\mathbf{x}_t + \mathbf{b}_{ke} \tag{1}$$

Figure-1 illustrates the model.

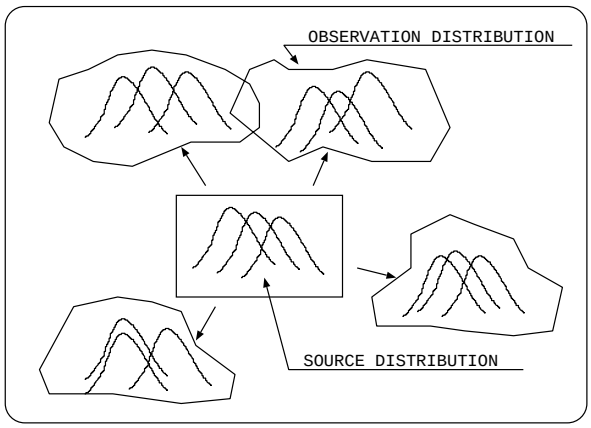

Figure 1: For a given sound (e.g.: phone), the distribution of the observations in different (class of) environments is transformed from a source distribution of the parameter space (e.g.: MFCC), by a linear transformation (shown as "$\rightarrow$") which is dependent on the environment, state and mixture component.

Notice that $\mathbf{x}_t$ is not observable and the distribution of $\mathbf{x}$ is not known, which differs the present work from model adaptation schemes, e.g. [5]. Also, in [2, 1], a supervision signal on the environment e must be given, e.g.: speaker identity, male/female. This work overcomes this potential limitation by allowing *unsupervised* training, i.e.: any desired number of environments can be specified which are optimally trained from the database. Therefore [2, 1] can be regarded as special cases of the formulations presented here.

Our problem now is to find, in the HMM framework and in the ML sense, the optimal source distributions, the transformation and the bias set.

Let N be the number of HMM states, M be the mixture number, L be the number of environments, $\Omega_s \triangleq \{1, 2, \ldots N\}$ be the set of states, $\Omega_m \triangleq \{1, 2, \ldots M\}$ be the set of mixture indicators, and $\Omega_e \triangleq \{1, 2, \ldots L\}$ be the set of environment indicators.

For an observed speech sequence of T vectors: $\mathbf{O} \triangleq \mathbf{o}_1^T \triangleq (\mathbf{o}_1, \mathbf{o}_2, \ldots \mathbf{o}_T)$, we introduce state sequence $\Theta \triangleq (\theta_0, \ldots \theta_T)$ where $\theta_t \in \Omega_s$, mixture indicator sequence $\Xi \triangleq (\xi_1, \ldots \xi_T)$ where $\xi_t \in \Omega_m$, and environment indicator sequence $\Phi \triangleq (\varphi_1, \ldots \varphi_T)$ where $\varphi_t \in \Omega_e$. They are all unobservable.

As a generalization of mixture CDHMM [4], the joint probability of $\mathbf{O}, \Theta, \Xi$ and Φ given model λ can be written as:

$$p(\mathbf{O}, \Theta, \Xi, \Phi|\lambda) = u_{\theta_1} \prod_{t=1}^{T} c_{\theta_t \xi_t} b_{\theta_t \xi_t \varphi}(\mathbf{o}_t) a_{\theta_t \theta_{t+1}} l_\varphi \tag{2}$$

where

$$b_{jke}(\mathbf{o}_t) \triangleq p(\mathbf{o}_t|\theta_t = j, \xi_t = k, \varphi = e, \lambda) \tag{3}$$
$$= N(\mathbf{o}_t; \mathbf{W}_{je}\boldsymbol{\mu}_{jk} + \mathbf{b}_{ke}, \Sigma_{jk}), \tag{4}$$

where $N(x; m, \sigma)$ stands for Gaussian distribution with mean vector m and covariance matrix σ, and

$$u_i \triangleq p(\theta_1 = i), \qquad a_{ij} \triangleq p(\theta_{t+1} = j|\theta_t = i) \tag{5}$$
$$c_{jk} \triangleq p(\xi_t = k|\theta_t = j, \lambda), \qquad l_e \triangleq p(\varphi = e|\lambda) \tag{6}$$

2.2. ML parameter estimation

The model parameters can be determined by applying generalized EM-procedure [3], in which two kinds of data are involved: observable $\mathbf{X}$ and non-observable (hidden variables) $\mathbf{Y}$. The set $\{\mathbf{X}, \mathbf{Y}\}$ is called complete data. The EM algorithm maximizes, w.r.t. a new parameter set λ, the mathematical expectation of the log-likelihood of the complete data, conditioned on the observed data $\mathbf{X}$, and for a value $\bar{\lambda} \in \Lambda$ of the parameter. The expectation is taken over the sample space of the unobservable data Y:

$$\mathcal{Q}(\lambda|\bar{\lambda}) \triangleq \mathbf{E}_{\mathbf{Y}} \left\{ \log p(\mathbf{X}, \mathbf{Y}|\lambda) | \mathbf{X}, \bar{\lambda} \right\} \tag{7}$$

An important property of the EM algorithm is that the log likelihood $p(\mathbf{X}|\lambda)$ is guaranteed to improve monotonically until it reaches a stationary point.

For our case, hidden variables are $Y \triangleq \{\Theta, \Xi, \Phi\}$ and observables are $X \triangleq \{\mathbf{O}\}$. To derive re-estimation equations, the forward-backward variable in CDHMM formulation [4] must be extended. Denote:

$$\alpha_t(j, e) \triangleq p(\mathbf{o}_1^t, \theta_t = j, \varphi = e|\bar{\lambda}) \tag{8}$$
$$\beta_t(j, e) \triangleq p(\mathbf{o}_{t+1}^T|\theta_t = j, \varphi = e, \bar{\lambda}) \tag{9}$$
$$\gamma_t(j, k, e) \triangleq p(\theta_t = j, \xi_t = k, \varphi = e|\mathbf{O}, \bar{\lambda}) \tag{10}$$

By equating the derivative of (7) w.r.t. each parameter of the model to zero, and solve the resulting joint equations, we can obtain the re-estimation formulae of all the parameters.

2.2.1. Initial state probability

$$u_i = \frac{1}{R} \sum_{r=1}^{R} \frac{\sum_{e \in \Omega_e} \alpha_1^r(i, e) \beta_1^r(i, e)}{\sum_{i \in \Omega_s} \sum_{e \in \Omega_e} \alpha_1^r(i, e) \beta_1^r(i, e)} \tag{11}$$

with R the number of training tokens.

2.2.2. Transition probability

$$a_{ij} = \frac{\bar{a}_{ij} \sum_{r=1}^{R} \frac{1}{p(\mathbf{O}^r|\bar{\lambda})} \sum_{e\in\Omega_e} \sum_{t=1}^{T^r} \alpha_t^r(i,e) b_{je}(\mathbf{o}_{t+1}^r) \beta_{t+1}^r(j,e)}{\sum_{r=1}^{R} \frac{1}{p(\mathbf{O}^r|\bar{\lambda})} \sum_{e\in\Omega_e} \sum_{t=1}^{T^r} \alpha_t^r(i,e)\beta_t^r(i,e)} \tag{12}$$

2.2.3. Mixture component probability

$$c_{jk} = \frac{\sum_{r=1}^{R} \sum_{e\in\Omega_e} \sum_{t=1}^{T^r} \gamma_t^r(j,k,e)}{\sum_{r=1}^{R} \frac{1}{p(\mathbf{O}^r|\bar{\lambda})} \sum_{e\in\Omega_e} \sum_{t=1}^{T^r} \alpha_t^r(j,e)\beta_t^r(j,e)} \tag{13}$$

2.2.4. Environment probability

$$l_e = \frac{1}{R} \sum_{r=1}^{R} \frac{\sum_{j\in\Omega_s} \alpha_T^r(j,e)}{\sum_{e\in\Omega_e}\sum_{j\in\Omega_s} \alpha_T^r(j,e)} \tag{14}$$

2.2.5. Mean vector and bias vector

We introduce:

$$\rho(j,k,e) \triangleq \sum_{r=1}^{R} \sum_{t=1}^{T^r} \gamma_t^r(j,k,e)\mathbf{o}_t^r \tag{15}$$

$$\varrho(j,k,e) \triangleq \sum_{r=1}^{R} \sum_{t=1}^{T^r} \gamma_t^r(j,k,e) \tag{16}$$

and

$$\mathbf{G}_{ke} = \sum_{j\in\Omega_s} \varrho(j,k,e)\Sigma_{jk}^{-1} \tag{17}$$

$$\mathbf{E}_{jke} = \varrho(j,k,e)\mathbf{W}_{je}{}'\Sigma_{jk}^{-1} \tag{18}$$

$$\mathbf{F}_{jk} = \sum_{e\in\Omega_e} \mathbf{E}_{jke}\mathbf{W}_{je} \tag{19}$$

$$\mathbf{a}_{jk} = \sum_{e\in\Omega_e} \mathbf{W}_{je}{}'\Sigma_{jk}^{-1}\rho(j,k,e) \tag{20}$$

$$\mathbf{c}_{ke} = \sum_{j\in\Omega_s} \Sigma_{jk}^{-1}\rho(j,k,e). \tag{21}$$

In the framework of generalized EM, we can assume $\mathbf{W}_{je} = \overline{\mathbf{W}_{je}}$ and $\Sigma_{jk}^{-1} = \overline{\Sigma_{jk}}^{-1}$, for a given k, we have $N+L$ equations:

$$\sum_{e\in\Omega_e} \mathbf{E}_{jke}\mathbf{b}_{ke} + \mathbf{F}_{jk} \qquad \mu_{jk} = \mathbf{a}_{jk} \quad \forall j \in \Omega_s \tag{22}$$

$$\mathbf{G}_{ke}\mathbf{b}_{ke} + \sum_{j\in\Omega_s} \mathbf{H}_{jke} \qquad \mu_{jk} = \mathbf{c}_{ke} \quad \forall e \in \Omega_e \tag{23}$$

Therefore μ_{jk} and $\mathbf{b}_{ke}$ can be simultaneously obtained by solving the linear system of $N+L$ variables.

2.2.6. Variance

$$\Sigma_{jk} = \frac{\sum_{e\in\Omega_e}\sum_{r=1}^{R}\sum_{t=1}^{T^r} \gamma_t^r(j,k,e)\delta_t^r(j,k,e)\delta_t^r(j,e,k)'}{\sum_{e\in\Omega_e} \varrho(j,k,e)} \tag{24}$$

where $\delta_t^r(j,k,e) \triangleq \mathbf{o}_t^r - \mathbf{W}_{je}\mu_{jk} - \mathbf{b}_{ke}$

2.2.7. Transformation

We assume covariance matrix to be diagonal: $\Sigma_{jk}^{-1(m,n)} = 0 \quad$ if $n \neq m$. Similar to [5], for the line m of $\mathbf{W}_{je}$, we can derive:

$$\mathbf{Z}_{je}^{(m)} = \mathbf{W}_{je}{}^{(m)}\mathbf{R}_{je}(m) \tag{25}$$

which is a linear system of D equations, where:

$$\mathbf{Z}_{je}^{(m,n)} \triangleq \sum_{k\in\Omega_m} \Sigma_{jk}^{-1(m,m)}\mu_{jk}{}^{(n)} \sum_{r=1}^{R}\sum_{t=1}^{T^r} \gamma_t^r(j,k,e)(\mathbf{o}_t^r - \mathbf{b}_{ke})^{(m)} \tag{26}$$

$$\mathbf{R}_{je}^{(p,n)}(m) \triangleq \sum_{k\in\Omega_m} \Sigma_{jk}^{-1(m,m)}\mu_{jk}{}^{(p)}\mu_{jk}{}^{(n)} \sum_{r=1}^{R}\sum_{t=1}^{T^r} \gamma_t^r(j,k,e). \tag{27}$$

If the means of the source distributions (μ_{jk}) are assumed constant (trained by conventional CDHMM reestimation procedure), then the above set of source normalization formulae can also be used for MLLR model adaptation [5].

3. EXPERIMENTAL EVALUATION

3.1. Database

The recognition task has 53 commands of 1-4 words. ("call return", "cancel call return", "selective call forwarding", etc). Utterances are recorded through telephone lines, with a diversity of microphones, including carbon, electret and cord-less microphones and hands-free speaker-phones. Some of the training utterances do not correspond to their transcriptions, due to mis-detection of speech/non-speech. For example: "call screen" (cancel call screen), "matic call back" (automatic call back), "call tra" (call tracing). No special treatment was devoted to these cases.

The speech is 8kHz sampled with 20ms frame rate. The observation vectors are composed of LPCC derived 13-MFCC plus regression based delta MFCC. CMN is performed at the utterance level for all tests. There are 3505 utterances for training and another 720 for speaker-independent testing. The number of utterances per call ranges between 5-30.

3.2. Experiments

Because of data sparseness, besides transformation sharing among states and mixtures, the transformations need to be shared by a group of phonetically similar phones. The grouping, based on an automatic hierarchical clustering of

phones, is dependent on the amount of data for training or for adaptation, i.e., the larger the number of tokens is, the larger the number of transformations. Each recognition experiment uses either of the following HMM training options:

- BASELINE uses conventional CDHMM. Notice that per-utterance cepstral mean normalization (CMN) is used (as in all other options). As simple technique, CMN will remove channel and some long-term speaker specificities, if the duration of the utterance is long enough, but cannot deal with time domain additive noises.
- SNCDHMM performs source-normalized CDHMM training as described in the paper, where the utterances of a phone call are assumed to have been generated by a call-dependent acoustic source. Speaker, channel and background noise that are specific to the call is reduced by SN-MLLR. We evaluated a special case, where each call is modeled by one environment. In recognition, only source distribution parameters are used and transformation and bias obtained during the training are discarded.
- BASELINE+AD adapts traditional HMM (BASELINE) parameters by unsupervised MLLR. 1. using current HMMs and task grammar to phonetically recognize the test utterances, 2. mapping the phone labels to a small number (N) of classes, which depends on the amount of data in the test utterances. 3. estimating the LR using the N-classes and associated test data. A similar procedure has been introduced in [5].
- SNCDHMM+AD refers to MLLR adaptation with seed models trained by SNCDHMM technique.

The resulting acoustic models are then used for recognition tests.

3.3. Findings

Based on the results summarized in Table-1, we point out:
A. For numbers of mixture components per state smaller than 16, SNCDHMM, BASELINE+AD, and SNCDHMM +AD all give consistent improvement over the baseline configuration.
B. For numbers of mixture components per state smaller than 16, SNCDHMM gives about 10% error reduction over the baseline. As SN is a training procedure which does not require any change to the recognizer, this error reduction mechanism can be immediately ported to applications.
C. For all tested configurations, MLLR adaptation using acoustic models trained with SN procedure always gives additional error reduction, compared to using models trained with conventional re-estimation procedures.
D. The most efficient case of SNCDHMM+AD is with 32 components per state, which reduces error rate by 23%, resulting 4.64% WER on the task.

4. CONCLUSION

We model speech phonetic variations by *source distributions* and environmental variations by a set of linear transformations and biases. The separation of the two types of random quantities makes it possible to share, among different environments, the information on the phonetic variations through source distributions.

We extended Baum's forward-backward variables to take into account of non-observable environmental variables and derived a set of re-estimation equations for source-normalized CDHMM. The unsupervised environment training allows any desired number of environments. By setting $W_e = 1$ and $b_e = 1$, the SN-CDHMM reduces to CDHMM. By fixing source distribution parameters, i.e., mean vectors and covariance matrices, the formulation can be used for MLLR model adaptation.

Experiments show that for the given database, MLLR adaptation from seed models provided by source normalized training procedure consistently gives lower word error rate than from models by conventional Baum-Welch procedure.

	4	8	16	32
BASELINE	7.85	6.94	6.83	5.98
SNCDHMM	7.53	6.35	6.51	6.03
BASELINE+AD	7.15	6.41	5.61	5.87
SNCDHMM+AD	6.99	6.03	5.41	4.64

Table 1: word error rate (%) as function of test configuration and number of mixture components per state

5. REFERENCES

[1] A. Acero and X. Huang. Speaker and gender normalization for continuous-density hidden Markov models. In *Proc. of IEEE Internat. Conf. on Acoustics, Speech and Signal Processing*, pages 342–345, Atlanta, 1996.

[2] T. Anastasakos, J. McDonough, R. Schwartz, and J. Makhoul. A compact model for speaker-adaptive training. In *Internat. Conf. on Spoken Language Processing*, volume 2, October 1996.

[3] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, 39(1):1–38, 1977.

[4] B. Juang. Maximum-likelihood estimation for mixture multivariate stochastic observations of Markov chains. *The Bell System Technical Journal*, pages 1235–1248, July-August 1985.

[5] C. J. Leggetter and P. C. Woodland. Maximum likelihood linear regression for speaker adaptation of continuous density HMMs. *Computer, Speech and Language*, 9(2):171–185, 1995.

[6] A. Sankar and C.-H. Lee. A maximum-likelihood approach to stochastic matching for robust speech recognition. *IEEE Trans. on Speech and Audio Processing*, 4(3):190–202, 1996.

[7] O. Siohan, Y. Gong, and J.-P. Haton. Comparative experiments of several adaptation approaches to noisy speech recognition using stochastic trajectory models. *Speech Communication*, 18:335–352, 1996.