論文 / 著書情報
Article / Book Information

Title	Robust Speech Recognition In The Car Environment
Author	Agnieszka Betkowska Cavalcante, Koichi Shinoda, Sadaoki Furui
Journal/Book name	the 4th Language and Technology Conference (LTC'09), Vol. , No. , pp. 39-43
発行日 / Issue date	2009, 11

Proceedings of the 4th Language and Technology Conference (LTC'09), Poznan, Poland, pp. 39-43, November 2009.

Robust Speech Recognition In The Car Environment

Agnieszka Betkowska Cavalcante*, Koichi Shinoda†, and Sadaoki Furui†

* Telcordia Poland
Umultowska 85, 61-614 Poznań, Poland
agabet@telcordia.com

† Tokyo Institute of Technology
2-12-1, Ookayama, Meguro-ku,Tokyo, 152-8552 Japan
{shinoda@, furui@furui}.cs.titech.ac.jp

Abstract

In this study we focus on robust speech recognition in car environments. For this purpose we used weighted finite-state transducers (WFSTs) because they provide an elegant, uniform, and flexible way of integrating various knowledge sources into a single search network. To improve the robustness of the WFST speech recognition system, we performed nonlinear spectral subtraction (SS) to suppress noise from the noisy speech. Using the "clean" speech signal obtained from SS, we conducted supervised WFST network adaptation to the characteristics of a given driver. In the best case, for highly noisy conditions, the speaker dependent WFST decoder achieved 70% improvement when compared with traditional speaker independent speech recognition systems.

Keywords: robust speech recognition, weighted finite-state transducers (WFST)

1. Introduction

Modern cars are being equipped with various electronic devices such as mobile phones, radios, and navigation systems. These give drivers more functionality and help, but their manipulation increase the risk of car accidents because they distract the drivers from the main task: driving. Therefore, a great deal of effort has been devoted to increasing safety by devising unobtrusive devices with hands-free interfaces.

Automatic speech recognition (ASR) technology is the most promising hands-free interfaces. However, this technology must be robust to be successful, i.e., it must work well in the presence of different noises appearing in car environments (street noise, engine noise, stray speech, etc), and it must be able to recognize speech of different speakers, each of whom has a unique speaking style and voice characteristics. Two common approaches to increase the robustness of ASR systems are the manipulation of the input signal, which removes unwanted noise (Gong, 1995), and optimization of a given speech model, which tries to improve the modeling of the signal for a given environment and speaker (Huang et al., 2001).

Current state-of-the-art ASR systems, frequently based on Hidden Markov Models (HMM) (Rabiner, 1989), are complex because they must deal with various knowledge sources at multiple levels, such as the acoustic model (the acoustic signal representation), the dictionary (transcription of words), and the language model (grammar representation), etc. Usually, each level of the ASR system is modeled and optimized separately because they have different model representation. Joint optimization of the overall system is difficult because the information about the relations between different sources is not available.

Recently, a new approach based on weighted finite-state transducers (WFSTs) (Caseiro and Trancoso, 2002) has received considerable attention from the speech community. WFST is a finite state network that encodes mappings between input and output symbol sequences (e.g., the mapping from an acoustic signal to a word sequence), and each mapping can be weighted with a log probability value. One of the main advantages of this approach as compared with traditional systems is that it is an elegant, uniform, and flexible way of integrating various knowledge sources into a single network (Mohri et al., 2002). Hence, optimizing the overall system is possible by adapting the parameters of a single WFST network instead of adapting the parameters of each knowledge source separately.

In this study we focus on robust speech recognition in car environments. We combine signal manipulation and model optimization techniques to increase the speech recognition performance for the individual drivers in the presence of engine/street noise. To improve the robustness of the speech recognition system, we performed nonlinear spectral subtraction (SS) to suppress noise from the noisy speech. Using "clean" speech signal obtained from SS, we conducted the supervised WFST network adaptation to the characteristics of a given driver. The evaluation was done with the "Driver's Japanese Speech Corpus in Car Environment', which was recorded as a part of the METI project, a project supported by the Japanese Ministry of Economy, Trade and Industry. The results of the initial experiments are promising. In the best case, for highly noisy conditions (signal to noise ratio (SNR) ranging from 0 to 8dB), the speaker dependent WFST decoder achieved 70% improvement when compared with traditional speaker independent speech recognition systems.

[0]This work was done while the first author was a project researcher with the Department of Computer Science, Tokyo Institute of Technology

2. Background

2.1. Spectral subtraction

In this study, we use two common assumptions in spectral subtraction: clean speech is corrupted by additive noise and both signals are uncorrelated. At each frame i, the clean speech power is estimated by subtracting the estimated noise power from the noisy speech power in the frequency domain, i.e.,

$$|\hat{X}_i(\omega)|^2 = \begin{cases} |Y_i(\omega)|^2 - \alpha_i|\hat{N}_i(\omega)|^2, \\ \quad \text{if} \quad |Y_i(\omega)|^2 > (\alpha_i + \beta)|\hat{N}_i(\omega)|^2 \\ \beta|\hat{N}_i(\omega)|^2, \quad \text{otherwise}, \end{cases} \tag{1}$$

where ω, $\hat{X}_i(\omega)$, $Y_i(\omega)$, $\hat{N}_i(\omega)$, and β are the angular frequency, the estimate of clean speech, the noisy speech, the estimate of the noise, and a spectral flooring factor, respectively. The parameter α_i, which has to be defined *a priori*, is a constant called the overestimation factor which depends on the signal to noise ratio (SNR) (Gong, 1995). The noise signal is unknown, so it has to be estimated by taking averages of frames that are known to be silence (Gong, 1995). The noise estimation is valid as long as the noise is stationary or its characteristics change relatively slowly compared to those of the speech signal.

2.2. Weighted finite state transducers in speech recognition

WFST is a finite state network that encodes mappings between input and output symbol sequences (for example, the mapping from an acoustic signal to a word sequence), and each mapping can be weighted with a log probability value. WFST algorithms include, among others, composition (combination of transducers), determinization, and minimization operations. Further details on the mathematical representation and algorithms can be found in (Mohri et al., 2002).

The application of WFST into a speech recognition system requires that each knowledge source is represented by a weighted transducer. These transducers are combined using the composition operation ($\circ$) into a single WFST network:

$$R = H \circ C \circ L \circ G, \tag{2}$$

where H, C, L and G are WFSTs representing the acoustic model (HMMs), the phoneme context-dependency, the dictionary and the language model, respectively. The network R can be optimized by performing determinization and minimization operations that lead to considerable improvement of the decoder's search efficiency (Mohri et al., 2002).

2.3. Adaptation of the acoustic model HMM

A common method for WFST network adaptation consists of two steps. First, adaptation of the acoustic and language models are performed separately. Then the adapted models are combined into a single WFST network. In this section we focus on the user adaptation of the acoustic model represented by HMMs.

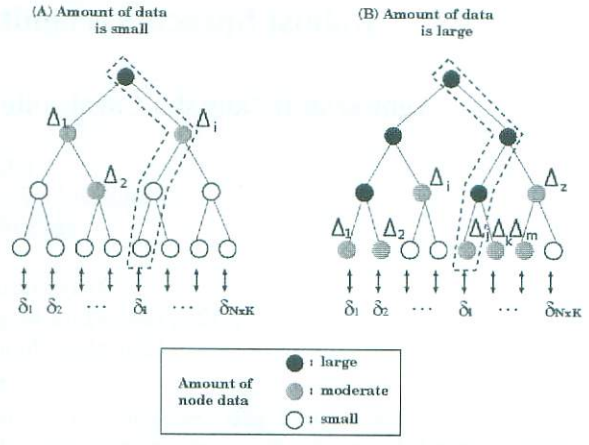

Figure 1: Tree structure for shifts estimation.

HMM is a set of N states, each of which can emit an output observation x (from a set of observations) with a given output probability density function (*pdf*). For each frame, let $\mathbf{x} = [x_1, x_2, \ldots, x_D]^T$ be the D-dimensional Mel Frequency Cepstral Coefficients (MFCC) vector for the speech signal. Than the output *pdf* $p_j(\mathbf{x})$ for state j ($j = 1, \ldots, N$) in the HMM is usually represented by a mixture of Gaussians (Huang et al., 2001):

$$p_j(\mathbf{x}) = \sum_{m=1}^{M} c_{jm} N(\mathbf{x}|\mu_{jm}, \Sigma_{jm}), \tag{3}$$

where M is the number of Gaussians in each state, μ_{jm} is the mean vector of the m-th mixture components of the state j, and c_{jm} is the m-th mixture coefficient, respectively. We assume that covariance matrix Σ_{jm} of the m-th mixture in state j is diagonal. The transitions among states are governed by the stochastic state transition matrix $A = \{a_{jk}\}$, where each element a_{jk} represents the probability of transition from state j to state k.

For an acoustic model defined this way, we use an adaptation method proposed by (Shinoda and Watanabe, 1995). In this method, the mean of each Gaussian component in the speaker-independent HMM (SI-HMM) is mapped to the unknown mean of the corresponding Gaussian component in the speaker-dependent HMM (HD-HMM). Let μ_i and $\hat{\mu}_i$ be the mean of the i-th Gaussian component of the SI-HMM and the corresponding Gaussian component of the SD-HMM, respectively. Then,

$$\hat{\mu}_i = \mu_i + \delta_i, \quad i = 1, \ldots, N \times M,$$

where δ_i is a shift parameter from the mean of the SI-HMM, N is the number of states in the model, and M is the number of Gaussian components in each state. The shift δ_i is estimated using a training algorithm such as the forward-backward algorithm or the Viterbi algorithm.

The number of required shifts δ_i is very large ($N \times M$) in general, so the correct estimation of these shifts with a limited amount of adaptation data is often very difficult. To overcome this problem, the proposed method controls the number of shifts to be estimated by using a tree structure of Gaussian components (see Figure 1). This tree is

constructed by clustering the Gaussian mixtures of all the states of the SI-HMM with a top-down clustering method that employs the k-means algorithm. The Kullback-Leibler divergence is used as a measure of distance between two Gaussians. In such a tree, each leaf node i corresponds to a Gaussian mixture i, and a tied-shift Δ_j is defined for each nonleaf node j. Using this tree structure, we control the number of free parameters according to the amount of data available. When we do not have a sufficient amount of data, a tied-shift Δ_j in the upper part of the tree is applied to all the Gaussian components below node j. As the amount of data increases, tied-shifts in the lower levels are chosen for adaptation. To control this process, we use a threshold that defines the minimum amount of data needed to estimate Δ_j. This threshold represents the number of data frames needed for the precise estimation of the shifts attached to each node and is chosen experimentally.

3. Experimental conditions

3.1. Drivers Japanese Speech Corpus in Car Environment

For the evaluation we used the "Driver's Japanese Speech Corpus in Car Environment" recorded by the Asahi Kasei company. The speech of each driver was recorded in the Toyota Vitz car with the microphone installed near the map lamp. For our study we used recordings of 110 nonprofessional female drivers, 110 nonprofessional male drivers, 20 professional female drivers, and 20 professional male drivers. The recordings were taken in three different conditions: with the car in an idle state, in a driving school (nonprofessional drivers) or in a city (professional drivers), and on highways (professional drivers only). The recordings taken on highways have low SNR, ranging from (0 to 8 dB).

For each condition each driver uttered around 270 450 commands. The commands included navigation commands, hands-free commands, and digit sequences, among others. Each command consists of several words.

The samples were digitized at 16 kHz sampling rate and analyzed with a frame of 10 msec. Mel Frequency Cepstral Coefficients consisting of 12 static features, 12 Δ features, 12$\Delta\Delta$ features, and Δ and $\Delta\Delta$ energy were used as an input vector in each frame.

3.2. WFST network construction

3.2.1. Acoustic model

We constructed a baseline speech left-to-right HMMs with 3 states and 16 Gaussian components per state. The basic recognition unit for the acoustic models were triphones, trained with a Japanese speech corpus for large vocabulary continuous speech recognition (LVCSR) systems called JNAS (JNAS,) (Japanese Newspaper Article Sentence). The constructed acoustic model was then converted to a WFST using AT&T FSM tools (Mohri et al., 1997)

3.2.2. Language model

The language model was constructed as follows. First, a grammar defined with extended Backus Naur Form (BNF) notation was parsed with the HTK HParse tool (Young et al., 2006), and a word network in a HTK standard lattice format (SLF) was generated. Next, the word network was converted to the WFST format. Finally a language WFST was compiled with the AT&T FSM library.

For each command group a separate language model was developed with a separate dictionary. The vocabulary list consisted of 286, 83, and 13 words for navigation commands, hands-free commands, and digit sequence task, respectively.

3.3. Evaluation of the baseline model

We built a separate baseline WFST recognition system for navigation, hands-free control, and digit sequence tasks. The resulting WFST decoders shared the same acoustic model, but their language model and dictionary were different. Each of these decoders was tested by professional and nonprofessional drivers' utterances in three conditions: idle state, city/school, and highway.

The system achieved the highest performance when the car was in an idle state. This result is expected because those test samples have high SNR. The recognition accuracy varied from 94.5% 98.6% for professional drivers and from 95.5% 99.1% for nonprofessional drivers, respectively. The worst results were achieved for a digit sequence task. It can be explained by the fact that the grammar is more flexible as compared to other tasks, so the recognition heavily relies on the accuracy of the acoustic model.

The WFST decoder achieved slightly worse results in city conditions.

The baseline WFST decoders failed in highway conditions, achieving recognition accuracy below 30% for professional female drivers and around 50% for professional male drivers. The low accuracy can be explained by a high level of noise due to the speed of the cars.

3.4. Evaluation of nonlinear spectral subtraction

We applied nonlinear spectral subtraction for speech recorded in a city and on a highway. When the car was in an idle state, the noise distortion of the speech was small, so there was no need for performing this procedure.

Proper estimation of the noise signal (based on silence frames) is essential for successful spectral subtraction. In the assumed scenario, the driver could speak anytime, so correct detection of silence and the speech part of the signal was crucial. To simplify the problem, the driver was requested to push a button before speaking. All frames of the signal recorded before the button was pushed were considered as silence. These frames were used for noise spectrum estimation. Spectral subtraction was applied to the signal taken after the button was pushed. The results of SS are shown in Figures 2 and 3.

A considerable improvement can be seen for both female and male professional drivers on highways. The best

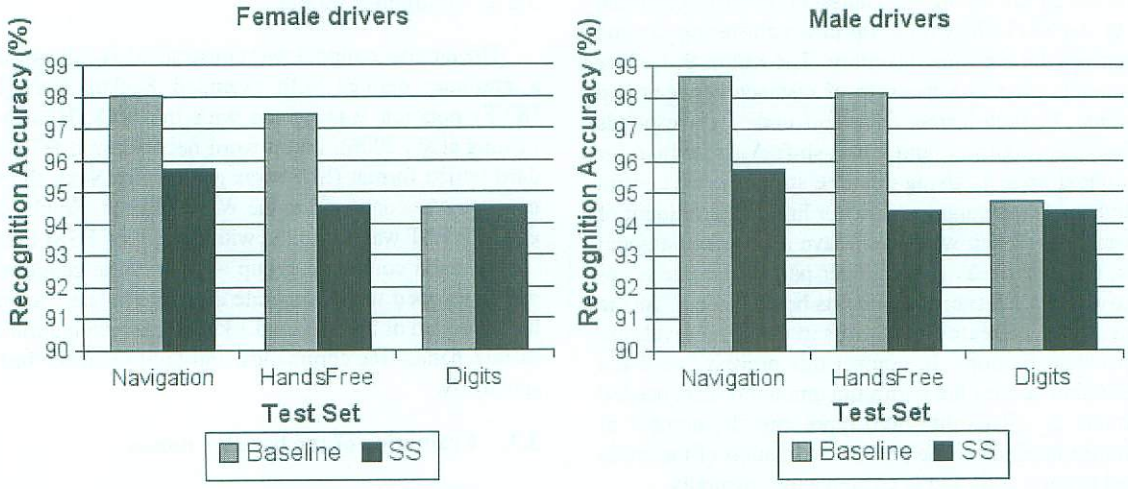

Figure 2: Spectral subtraction - nonprofessional drivers, city

results were achieved for the hands-free commands: 24.4% and 22.3% of absolute gain in accuracy for female and male speakers, respectively.

In the city, application of SS caused a slight degradation of the recognition performance of most of the drivers. Taking into account the high recognition performance of the baseline WFST decoder, we can assume that only a small amount of noise is corrupting the speech. Distortions caused by SS (so called musical noise) seem to give more problems for the recognition system than the noise originally corrupting the speech signal.

3.5. Speaker adaptation of acoustic model

We performed speaker adaptation for professional drivers driving on highways. Each speaker data was randomly divided in half to create the adaptation and test sets. For the speaker adaptation, we used Shinoda's algorithm. The tree of the Gaussian components had four levels and four branches at each level. The threshold, which defines the minimum frames needed for the tied-shift Δ estimation, was experimentally set to four. Before adaptation, spectral subtraction was applied to the adaptation and test sets. We created a speaker-dependent acoustic model for each driver by adapting a baseline (speaker-independent) acoustic model. The speaker-dependent acoustic models were combined later with the language model to create a single WFST network. We tested the adapted WFST decoder for each speaker and for each command task. The results are shown in Figure 3.

Compared to the baseline WFST decoder, the speaker adaptation improved the WFST decoder's performance in all three tasks. In case of the digits task the improvement was 49.4% and 29.9% absolute for professional female drivers and professional male drivers, respectively. In the hands-free task, the improvement was 70.6% and 46.6% absolute. The best results, 93.5% and 94% recognition accuracy, were obtained for the navigation command task. The digit sequence task, as in the baseline WFST decoder, was the most difficult to be recognized correctly (e.g. 78% of recognition accuracy for female drivers).

4. Conclusions

We investigated different methods for improving the speech recognition performance in car environments. The discussed methods were evaluated with the "Driver's Japanese Speech Corpus in Car Environment" and compared against a baseline WFST speech recognition system.

The baseline WFST system achieved high recognition accuracy (over 95%) for the samples recorded in a car in the idle state. Slightly worse performance was obtained for samples recorded in the city. A significant degradation of the recognition performance was observed in highways. However, applying nonlinear spectral subtraction (SS) in highway environment improved the recognition performance of the WFST network by 24.4% and 22.3% absolute for professional female and male speakers, respectively.

We also evaluated the WFST system optimized in highway conditions. The combined optimization method included spectral subtraction and speaker adaptation. In the best case, for professional female drivers with hands-free command task, the performance of the WFST network was increased by 70.6% absolute.

In this study, the WFST adaptation is done in two steps. First, the acoustic model is adapted to the characteristics of the specific driver. Then this model is combined with language model into a single WFST network. Unfortunately, this scheme for adaptation does not take into consideration the relation between acoustic and language models, and it can only be performed offline. In the future, the supervised and unsupervised adaptation of the whole WFST network should be investigated and compared with traditional optimization methods.

Acknowledgments

This works was supported by the Japanese government METI Project "Development of Fundamental Speech

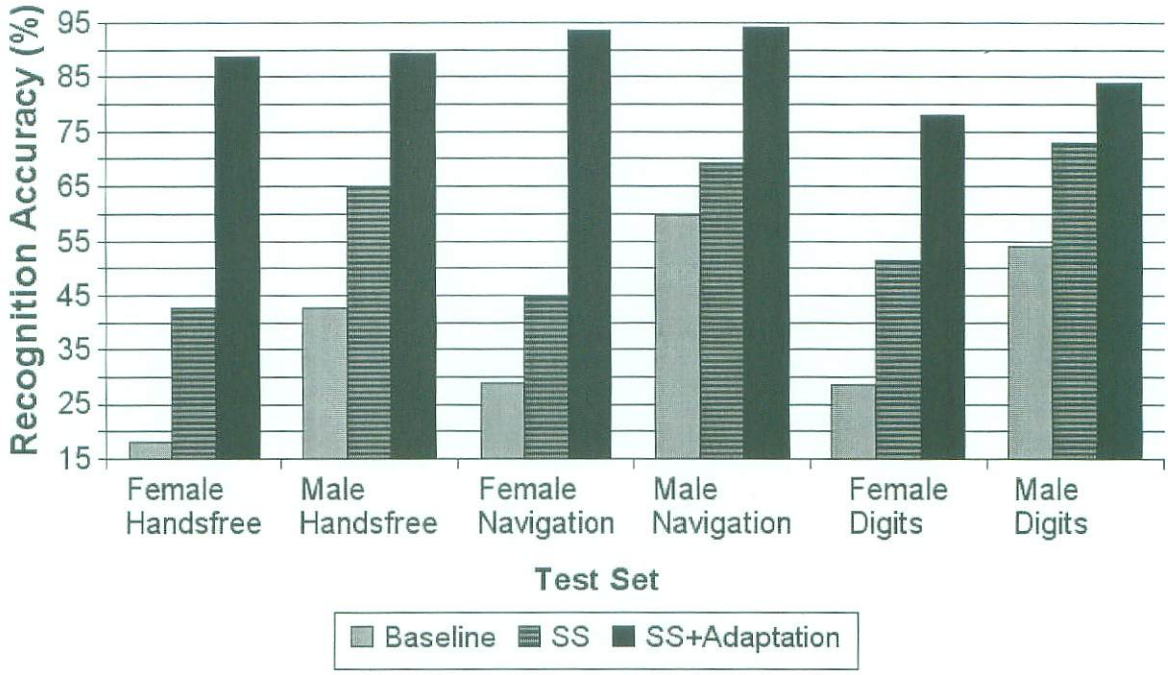

Figure 3: Adaptation of WFST network and SS - Professional drivers, highway

Recognition Technology".

References

Caseiro, D. and I. Trancoso, 2002. Using dynamic wfst composition for recognizing broadcast news. In *International Conference on Spoken Language Processing*.

Gong, Y., 1995. Speech recognition in noisy environments: A survey. *Speech Communication*, 16(3):261–291.

Huang, X., A. Acero, and H. Hon, 2001. *Spoken Language Processing: A Guide to Theory, Algorithm and System Development*. USA: Prentice Hall.

JNAS. : Japanese newspaper article sentences, http://www.mibel.cs.tsukuba.ac.jp/jnas/.

Mohri, M., F. C. N. Pereira, and M. Riley, 1997. A rational design for a weighted finite-state transducer library. In *Lecture Notes in Computer Science*, volume 1436. pages 144 – 158.

Mohri, Mehryar, Fernando C. N. Pereira, and Michael Riley, 2002. Weighted finite-state transducers in speech recognition. *Computer Speech and Language*, 16(1):69–88.

Rabiner, R., 1989. A tutorial on hidden markov models and selected applications in speech recognition. In *Proceedings of the IEEE*, number 2.

Shinoda, K. and T. Watanabe, 1995. Speaker adaptation with autonomous control using tree structure. In *EuroSpeech95*.

Young, S. et al., 2006. *The HTK Book: HTK Tools and Reference Manuals, Version 3.4*. UK: Cambridge University Press.