INTERNATIONAL COMPUTER SCIENCE INSTITUTE

1947 Center St. • Suite 600 • Berkeley, California 94704-1198 • (510) 666-2900 • FAX (510) 666-2956

SPEECH RECOGNITION EXPERIMENTS ON SWITCHBOARD CORPUS

Toshihiko Abe [1]

TR-00-004

March 2000

Abstract

This report shows results of a set of speech recognition approaches performed on Switchboard corpus which is a large spontaneous telephone conversation database. The purpose is to improve recognition accuracy on Switchboard with our connectionist hybrid model. The methods include the choice of kinds of acoustic features, gender dependent training, use of multi-stream features, etc. We will also show that adding a feature of periodicity measure improves recognition accuracy. Finally, we will show a speaker adaptation approach that improves recognition accuracy for speech by a particular speaker.

[1] Research supported by the Japan Society for the Promotion of Science.

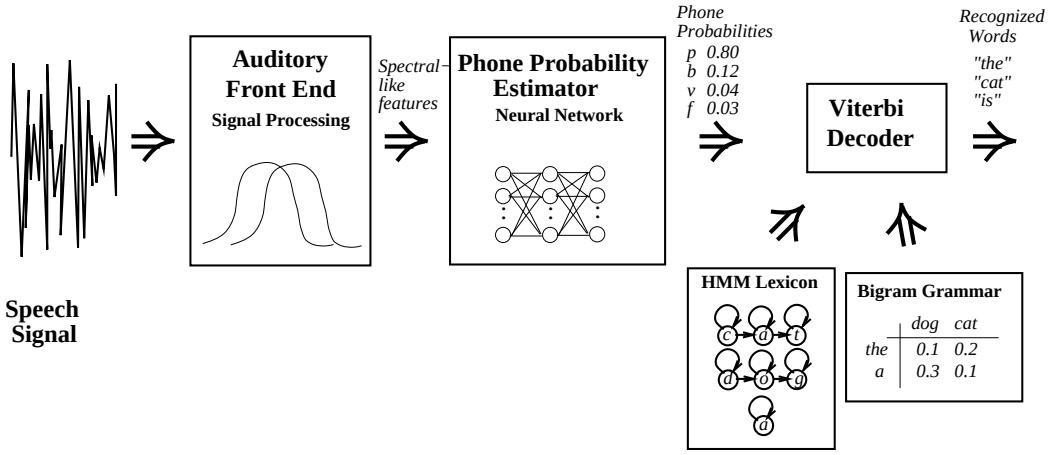

Figure 1.1: hybrid ANN-HMM system

1 INTRODUCTION

Automatic speech recognition has been a major goal for a large research community in the last few decades. The predominant approach to large vocabulary, speaker-independent, continuous speech recognition has been based on hidden Markov models(HMMs). Connectionist models have been widely proposed as a potentially powerful approach to speech recognition. Despite the very good results achieved in static pattern classification, there is not yet enough knowledge to adequately model the temporal structure of speech through connectionist models. To overcome these difficulties hybrid connectionist-HMM models have been tested. The connectionist system acts as a phone probability estimator and is used as the observation model within the HMM framework. This approach brings some benefits[1]. In particular, strong assumptions about the input statistics and the functional form of the output density are not required.

In this report, we describe a set of experiments designed to obtain good recognition for the Switchboard corpus[5] which is a human-to-human telephone conversation speech database, by incorporating existing various approaches that have been proven in their effectiveness. Those methods include the choice of acoustic features such as RASTA[3], PLP[2] and Modspec[6] and gender dependent training, use of multi-stream features, use of a feature of periodicity, speaker adaptation techniques, etc.

The spontaneous nature of speech in Switchboard makes the recognition job very challenging both in terms of acoustics and language. Many evaluations of large vocabulary speech recognition technology are currently done on Switchboard.

2 HYBRID ANN-HMM SYSTEM

Throughout this work, we use a hybrid Hidden Markov Model(HMM) and Artificial Neural Network(ANN) system for speech recognition[1](figure 1.1). In this system, the ANN estimates phone probabilities given the acoustic signal. The HMM is used to model the temporal structure of speech.

The ANN we use is a multilayer perceptron(MLP) that has a single hidden layer and local acoustic context through a multi-frame input window(figure 1.2). It estimates context-independent posterior phone probabilities. Gradient decent training is used to minimize the classification error via a cross-entropy criterion on the training utterances. Cross validation(CV) and early stopping are used to insure that the network parameters do not overtrain to new data. Details of the architecture and the training algorithm is described in [1].

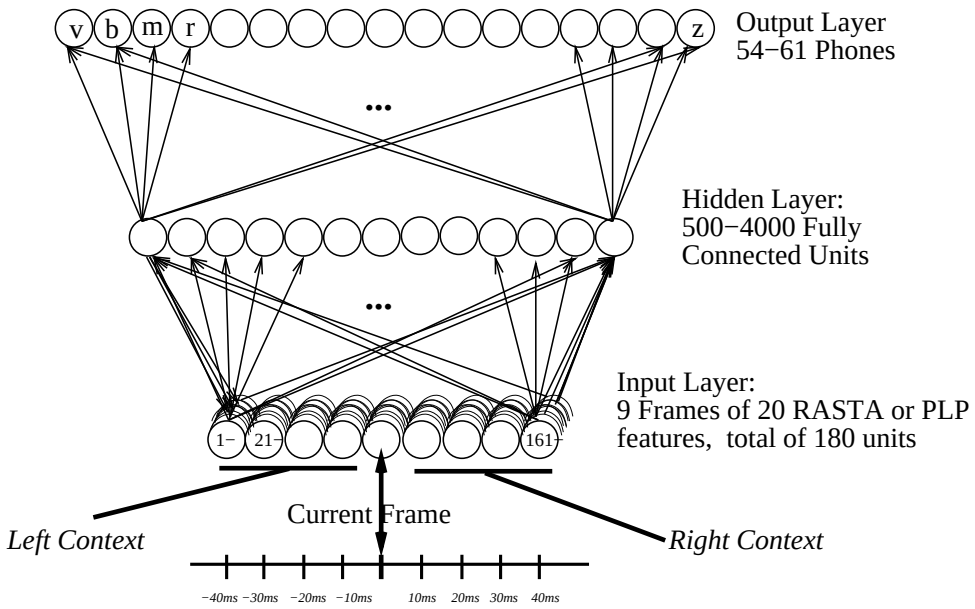

Figure 1.2: multilayer perceptron

3 HISTORY OF IMPROVEMENT

3.1 Baseline System

Table 1.1 shows the history of improvement of speech recognition of the ICSI hybrid system on Switchboard corpus. The first experiment(No.1) used the RASTA-PLP feature as a starting point. The Perceptual Linear predictive(PLP) feature estimates the auditory spectrum using approximations to the psychophysics of hearing[2]. RASTA stands for relative spectral analysis, which suppresses the components of spectral trajectories that change more slowly or more quickly than the statistically observed behavior of speech[3]. In the table 1.1, 'rasta12' means that there are 12 RASTA features. The leftmost column shows the word error rate(W.E.R.) As you can see, this baseline system gives 65.2% W.E.R.

3.2 Training Paradigm

The whole training set is 49 hours of Switchboard data selected by John Hopkins for 1996 system[5]. The training set is diveded into training(90%) and CV(10%) sets for the acoustic training. To expedite testing, the CV set was used for recognition; however the CV data was included in the language model training, so numbers of W.E.R. in the table1.1 are optimistic. The 'train set' column shows how much training data is used(in this case, half of speech data of 49 hours long). The MLP weights were initialized in a random fashion.

There are two iterations in each embedded training. Each iteration of the embedded training consists of 1) making the alignment of phone labels by the MLP trained in the previous iteration (using the boot alignment replaces that for the first iteration) and 2) MLP training. In general, there are 8 to 10 epochs in the MLP training for each iteration.

We also tried using boot nets as the initial parameters of the network. Boot nets means MLP weights trained on another corpus. In this case, they were trained on Switchboard and NTMIT databases. Boot nets are supposed to speed up training, but in this case they did not make any difference.

Table1.1: history of improvement

No.	method	net size	train set	test set	W.E.R.(%)
1	rasta12	13x9:2000:56	12.25h	cv300 (25min)	65.2
2	plp12norm	"	"	"	63.4
3	gender dependent training plp12norm	"	24.5h	"	59.5
4a	gender dependent training plp12norm 17-frame context window	"	"	"	56.5
4b	"	"	"	cv600 (50min)	57.2
5	gender dependent training per-side norm plp 17-frame context window	"	"	"	55.1
6	gender dependent training per-side norm modspec 17-frame context window	26x17:2000:56	"	"	57.9
7	combination of No.5 and No.6	13x17:2000:56 + 26x17:2000:56	"	"	51.8
8	twice the data twice the netsize	13x17:4000:56 + 26x17:4000:56	49h	"	49.6
9	new alignment new decoder	"	"	"	46.7

3.3 Normalized PLP features

The second experiment(No.2,'plp12norm') used per-utterance normalized PLP features. It has been reported that RASTA works better in general for relatively small recognition tasks, but per-utterance normalized plp can work better for large vocabulary tasks such as Switchboard. In this case, the PLP features were normalized for each utterance, instead of incorporating RASTA. The purpose is to get rid of static environment characteristics such as the channel effects. The result in W.E.R. 63.4% indicates slight improvement in comparison to using the RASTA-PLP, which gives 65.2% W.E.R.

Here again we tried using the pre-trained boot nets, but it did not make a difference. Since the pre-trained boot net does not seem to have any advantage, the random weight initialization will be adopted for the rest of experiments in this report.

3.4 Gender Dependent Training

The experiment No.3 incorporated the gender dependent(GD) training[4] to the previous(No.2) method. This time two nets were trained. The male net was trained with only male speech, and the female net was trained with only female speech.

Male and female speech are in general acoustically different, because of the difference in vocal length, pitch etc. Therefore using two separate models for them is better than using a single model. This is done by training two nets, one is trained with only male speech and the other with female speech. This framework is called GD training, while training a single net for both genders is called gender independent(GI) training.

In order to recognize speech whose gender is not known, two different methods for GD recognition are considered. One is training another net that estimates the correct gender and choose the GD

net of the correct gender for recognition. Another one is to first decode the speech with both GD nets, get the outputs with likelihood measure, and choose the more likely one as the final result.

The baseline score without GD training is 63.4% in W.E.R. that is shown in No.2 in table 1.1. The result on GD training for the first method was 59.3% W.E.R. with accuracy of 90.3% of the gender detection net. For gender detection, the gender detection MLP was trained with the same features as are used for recognition, with 2000 hidden units and two(which is the number of genders) output units. The second method gave 59.5% W.E.R which is shown in No.3 in table 1.1. Since those two numbers in W.E.R. are very close and the second method does not need a gender detection net and therefore is easier to implement, we choose the second as gender dependent training for further experiments in the rest of this report. To see the baseline, we also tried GD recognition by giving the correct gender to the system, instead of performing gender detection. The result was 59.8% W.E.R. which is worse than other systems. It is probably because there are some speakers whose speech is more similar to the opposite gender's speech and the gender selection system chose the opposite gender for those cases.

3.5 Longer Context Window VS Adding Delta Features

The experiment No.4 used a longer context window of 17 frames, instead of 9-frame window that had been used in previous experiments. Other than that, the same methods were used as the experiment No.3. For the experiment 4b, twice the size of the test data (CV 600 utterances) was used, in order to compare in subsequent experiments. Those 600 utterances were chosen to be gender-balanced(300male+300female).

Another option is to use delta features instead of a longer context window. In order to know which worked better, the same experiment except for this difference was performed. The result was 57.6% W.E.R., which did not exceed the one with the longer context window which gave 56.5%. Therefore using the longer context window was chosen for further experiments in the remainder of this report.

3.6 Per-side normalization of PLP features

The experiment No.5 uses per-side normalization of acoustic features, where the acoustic features are normalized over the conversation side, instead of over the utterance which had been used in the previous experiments. It suppresses static characteristics that are specific to the particular conversation such as the channel or environmental characteristics. The previous per-utterance normalization has also this function, but each utterance is short (about 5 seconds long in average), therefore per-side normalization is considered to be more suitable for this purpose. The result in W.E.R. is 55.1%.

Up to this, there have been two iterations in each embedded training, but it is found that training only once(no iterations) gives essentially the same W.E.R. result. Therefore in order to save training time, no iterations will be done for further experiments in this report.

3.7 Modulation Spectrogram

The experiment No.6 uses the modulation spectrogram (modspec) features instead of PLP. The primary hypothesis behind the modspec is that phonetic information is encoded in the speech signal as relatively slow changes in the spectral structure of speech[6]. Such a hypothesis matches the timing properties of the articulators and auditory cortical neuron activity. The modspec represents the speech signal as a distribution of slow modulations, from 0 to 8 Hz with a peak at 4 Hz, across time and frequency. The 4-Hz sensitivity corresponds roughly to syllabic frequencies. It has been found that the modspec performs slightly worse than PLP, but better for reverberant speech.

Except for this, the same methods were used as the experiment No.6, including the per-side normalization of the Modspec feature and the gender dependent training. The result in W.E.R. is 57.9%.

3.8 Combination of PLP and Modspec

The experiment No.7 is combination of the experiments No.5 and No.6. We have the nets which were trained on per-side normalized PLP and per-side normalized Modspec respectively. The probability streams that are the outputs of the PLP and Modspec nets were combined at the frame level. The combination was done by multiplying two probabilities. The result in W.E.R. is 51.8%, which means that the combination of the two different features works a lot better than either one of them alone.

3.9 Training with More Data

In the experiment No.8, twice the amount of training data was used. Also, twice the number of hidden units were used in accordance with the increase of the training data. Other than that, the same methods and features were used as No.7. The result in W.E.R. is 49.6%.

3.10 New Decoder

The decoder called 'noway' had been used in previous experiments. This was replaced with a newer decoder 'chronos', which often gives better performance than 'noway'.

Thus in the experiment No.9, the old decoder was replaced with the new decoder. It was used to generate phone labels by forced Viterbi alignment for training, and also used for recognition of the test set. Other than that, the same methods and features were used as No.7. The result is 46.7% in W.E.R and this is the best score for the CV set. However, since the CV set that was used for evaluation but also included in the language model training, the W.E.R of 46.7% is not really a fair performance measure. In order to get a more objective number of W.E.R., we will use the development set which is completely separate from any kind of training, in the next experiment.

3.11 ADDING PERIODICITY MEASURE TO ACOUSTIC FEATURES

3.11.1 Introduction

This section describes experiments of adding a new feature which indicates the periodicity of speech. Periodicity in a particular frame of speech indicates whether it is voiced or unvoiced. Therefore we can expect the periodicity measure will help discrimination of the voice and unvoiced phones.

3.11.2 Feature Extraction

The instantaneous frequency(IF) is defined as the angular velocity of an analytic signal. The definition of the periodicity measure in this section is based on the instantaneous frequency(IF) spectrogram, which is known to well describe the harmonic structure of speech[9]. This property makes it easy to define the periodicity measure based on the spectrum.

The periodicity measure $p(t)$ at time t is defined by

$$p(t) = \frac{\int_{\omega_0}^{\omega_1} \|F(\omega,t)\|^2 \left[1 - \exp\left\{-\frac{(m(\omega,t))^2}{\rho^2}\right\}\right] d\omega}{\int_{\omega_0}^{\omega_1} \|F(\omega,t)\|^2 d\omega + \alpha} \tag{1}$$

where $\|F(\omega,t)\|$ is the short time Fourier transform(STFT) amplitude spectrum and $m(\omega,t)$ is the normalized local second-order moment of the IF amplitude spectrum defined in [8]. The constant α is suppose to be a relatively small number so that the denominator will not be close to zero, which makes $p(t)$ abnormally large. The function $p(t)$ actually corresponds to aperiodicity, but we call it 'periodicity measure' for simplicity, since it does not make a difference when it is used as the input feature for training. The term $\exp\left\{-\frac{(m(\omega,t))^2}{\rho^2}\right\}$ corresponds to periodicity of the frequency component at ω. Therefore the term $\left[1 - \exp\left\{-\frac{(m(\omega,t))^2}{\rho^2}\right\}\right]$ corresponds to aperiodicity of the

Table 5.1: result on Numbers95

system	net size	W.E.R.(%)
rasta8+delta(baseline)	18x9:400:56	8.2
rasta8+delta+new feature	19x9:400:56	7.0

Table 5.2: result on dev-set from Switchboard

system	net size	W.E.R.(%)
plp+mod (baseline)	13x17:4000:56 + 26x17:4000:56	54.3
plp+mod+new feature	14x17:4000:56 + 27x17:4000:56	53.7

frequency component at ω, which takes values from 0 to 1. Then the integration of it over the frequency corresponds to the measure of aperiodicity at time t. The constant ρ determines the mapping of largeness of the moment and the degree of the periodicity for the frequency component. The constant ρ has been adjusted experimentally so that $p(t)$ visually shows good correspondence with voiced and unvoiced part of speech. Thus it is set to be $(11.05\text{Hz})^2$. The frequencies ω_0 and ω_1 are set to be 62.5Hz and 4000Hz, respectively. Values of $p(t)$ are not too sensitive to those numbers. Those fractional numbers are nothing special but because of the conversion from different units.

3.11.3 Experiments

As a preliminary experiment, first we tried adding the new feature $p(t)$ to a smaller task that is called Numbers95. It is a speech database of connected digits such as 'one hundred and thirty one.' The training set of 3590 utterances which includes 357 utterances for CV, and the test set of 1227 utterances are used. The RASTA features of ninth order and their deltas are used for the baseline system. The periodicity measure $p(t)$ is added in the same frame step rate as the RASTA features. The baseline system use 18 acoustic features and since the periodicity measure is a single scalar feature, the new system has 19 input features.

The results are shown in table 5.1. The baseline system gives 8.2% in W.E.R and the new system gives 7.0%. Thus improvement by absolute 1.2%, which is significant at 0.01 level, is obtained by adding the new feature.

We also tried on the development set of Switchboard. The development set has 2119 utterances. The baseline system is the one in the experiment No. 9 in table 1.1, which has shown the best score. The W.E.R. of the baseline system on the development set is 54.3% (table 5.2), which is quite higher than the W.E.R. in table 1.1. This is because the language model was trained on the training set and the cv600 set tested in table 1.1 is included in it, while the development set is not included in the language model training or any other kind of training. Therefore the score shown in table 5.2 is the most fair one. For the new system, the periodicity measure is given in parallel to both of the PLP and Modspec nets as an extra acoustic feature. Although this is redundant because the same information is given to two different nets, it keeps a kind of symmetry of the two nets in terms of given acoustic features.

The W.E.R. of the new system is 53.7%, therefore the improvement by adding the periodicity measure is absolute 0.6%. This may not seem a big improvement. However, considering that it is done by just adding one single feature, we can say that it is an appreciable improvement and that the periodicity measure is an effective acoustic feature in speech recognition.

Table 4.1: word error rate for speaker adaptation

speaker	gender	before adaptation	after adaptation	absolute improvement
2394B	male	41.9%	40.9%	1.0%
2800A	male	52.1	48.1	4.0
3056B	male	41.8	41.9	-0.1
2536B	male	46.1	42.6	3.5
2626A	male	40.0	38.1	1.9
2830B	female	59.9	57.8	2.1
2868A	female	41.4	40.2	1.2
2870A	female	44.9	45.2	-0.3
3092A	female	32.0	29.7	2.3
3151A	female	31.8	32.1	-0.3

Table 4.2: average word error rate for speaker adaptation

	before adaptation	after adaptation	absolute improvement
average on male	44.38%	42.32%	2.06%
average on female	42.00	41.00	1.00
total average	43.19	41.66	1.53

3.12 Conclusion

As is shown in the previous subsection, our best score is 53.7% W.E.R., which is better than the baseline experiment by almost 20%, considering that W.E.R. is lower by 10% for the CV set. As you can see in the table1.1, there is no method which contributes overwhelmingly, but each of the methods contributes to moderate amount of improvement. The number of 53.7% is not consider to be our final point. There are still more methods that can improve recognition accuracy, such as context dependent training, soft target training, syllabic models, new language model, new acoustic features etc.

4 SPEAKER ADAPTATION

4.1 Introduction

This section presents the results for speaker adaptation with the hybrid HMM-ANN system. This section is separate from the previous one, since the framework of the task is different in that 1) it is assumed that there are set of speakers each of which has certain amount of speech, 2) each speaker's speech is separate from others and 3) for any speech, identification of the speaker is given to the system in prior.

The speaker adaptation method is applied off-line to the speaker independent(SI) system to adapt the MLP to a new speaker. In all of the experiments, supervised adaptation is performed.

The method we use here for speaker adaptation is called Retrained Speaker-Independent(RSI)[7]. Starting from the SI system, all the net parameters are adapted to the new speaker. Adaptation is basically done by the same way as the SI system was trained, except for with much smaller rate of updating parameters in order to prevent overfitting of the net to the new speech.

4.2 Experiment of Speaker Adaptation

The experiment was performed with 10 speakers that consists of 5 male and 5 female speakers. For each speaker, 100 utterances were picked up and divided into two parts that are 75 utterances for

adaptation and 25 for CV and also W.E.R. evaluation. The jack knife method was adopted for evaluation, in which 4 different combinations for adaptation and CV are taken in rotation. In other words, either the first, second, third or the fourth quarter of the 100 utterances is used for CV and and the rest of 75 utterances are used for adaptation. For each speaker, either the male or female SI net is chosen for the initial net parameters before adaptation, according to the correct gender. This means that for training, we suppose not only the correct transcription but also the correct gender is known before adaptation. We use the system in the experiment No.9 as the speaker independent system for adaptation, since it shows the best score.

Table 4.1 shows the result. The total average of improvement in W.E.R. is 1.53%, which is significant at 0.05 level. This amount of 1.53% is about the same as the previous RSI result done in[7], where the improvement was 8.4%(SI) in W.E.R. to 6.9%(RSI), though the absolute value of W.E.R is much smaller because of the different speech corpus. The result in table 4.1 indicates that the adaptation works for most of speakers, but the improvement depends on the speaker. Since the amount of improvement varies a lot from speaker to speaker, 10 speakers may not be enough to see the average. Besides, we found no strong relationship between the absolute error rate before adaptation and the amount of improvement.

We also tried unsupervised training, where no transcription is given for the adaptation data. The adaptation data is given to the SI system and it generates transcription. Then it is used for alignment of phone labels. After speaker adaptation is done with the phone labels, the adaptation data is again given to the adapted recognition system, since we expect the improvement of recognition accuracy. Thus this forms a framework of iterations of recognition and training.

This unsupervised training is better if it works, since the need of preparing transcription limits possible applications. Unfortunately, trial experiments indicate no improvement. It seems that it is because of the transcription errors that were made in the recognition of the adaptive data by the SI system. However, it is expected that there still is possibility that unsupervised training will work, if the original SI system is good enough.

4.3 Discussion

It is shown that the speaker adaptation works for improvement in recognition accuracy. The amount of improvement is also moderate, just like the experiments in section 3. In this section, the RSI method is investigated. However, there may be other methods for speaker adaptation that will work better. The Linear Input Network(LIN) is such an example[7]. We also tried LIN, but we have not been able to achieve experiments, mainly because it needs a different network structure and a different training algorithm. However, it will be soon possible to use a software which has flexibility to implement any structure of ANN.

5 OVERALL CONCLUSION

For acoustic features, the combination of the per-side normalized PLP and the per-side normalized Modspec gives the best score. We have also shown that gender dependent training improves recognition accuracy. Training strategies such as increasing the amount of training data with more number of hidden units, and using the 'chronos' decoder instead of 'noway' also work for improvement. Adding a feature of periodicity also improves the recognition accuracy. Finally, we have shown that the speaker adaptation method improves recognition accuracy for speech by a particular speaker.

Our best score is 53.7% W.E.R., which is better than the baseline experiment by almost 20%, considering that W.E.R. is lower by 10% for the CV set. For future work, prospective methods that are expected to improve recognition accuracy further more include context dependent training, soft target training, syllabic models, new language model, new acoustic features etc.

Acknowledgements

I acknowledge Prof. Nelson Morgan for supervising my work and Eric Fossler for great amount of help in achieving experiments. Thanks to Dan Ellis, David Johnson and Jane Edwards for help in computer resources. And Thanks to Su-Lin Wu and Adam Janin for general help. Thanks also to Warner Warren, Mike Shire, Jeff Bilmes, Barry Chen and Shawn Chang for helpful advices and discussions.

Funding for my work is provided by the Japan Society for the Promotion of Science. Additional funding is from ICSI.

References

[1] Bourlard, Hervé and Nelson Morgan, *Connectionist Speech Recognition- A Hybrid Approach*, Kluwer Academic Press, 1994.

[2] Hermansky, Hynek, "Perceptual linear predictive (PLP) analysis of speech," *Journal of the Acoustical Society of America*, 87.1738–1752, 1990.

[3] Hermansky, Hynek and Nelson Morgan, "RASTA processing of speech," *IEEE Transactions on Speech and Audio Processing*, 2.578–589, 1994.

[4] Victor Abrash, Horacio Franco, Michael Cohen, Nelson Morgan, Yochai Konig, "Connectionist gender adaptation in a hybrid neural network / hidden Markov model speech recognition system," *Proceeedings of International Conference on Spoken Language Processing*, 1992.

[5] John Eric Fossler-Lussier, "Dynamic Pronunciation Models for Automatic Speech Recognition," Ph.D. thesis, University of California, Berkeley, 1999. Reprinted as ICSI technical report tr-99-015.

[6] Steven Greenberg and Brian E. D. Kingsbury, "The Modulation Spectrogram: In Pursuit of an Invariant Representation of Speech," *Proceedings of ICASSP*, vol. 3, pp. 1647–1650, 1997.

[7] J. Neto, L. Almeida, M. Hochberg, C. Martins, L. Nunes, S. Renals, T. Robinson, "Speaker-adaptation for hybrid HMM-ANN continuous speech recognition system", Eurospeech'95, 1995.

[8] T. Abe, T. Kobayashi, S. Imai, "Robust pitch estimation with harmonics enhancement in noisy environments based on instantaneous frequency," *Proc. ICSLP 96*, pp. 1277–1280, 1996.

[9] T. Abe, T. Kobayashi, S. Imai, "The IF spectrogram: a new spectral representation," *Proc. ASVA 97*, pp. 423–430, 1997.