The 1998 BBN BYBLOS Primary System applied to English and Spanish Broadcast News Transcription

Spyros Matsoukas, Long Nguyen, Jason Davenport, Jay Billa, Fred Richardson, Manhung Siu, Daben Liu, Rich Schwartz, John Makhoul

BBN Technologies
Cambridge MA 02138

ABSTRACT

In this paper, we describe the BBN BYBLOS system used for the 1998 Hub-4E primary and Hub-4Sp evaluation benchmarks, and discuss the improvements made to the system in 1998. We focus on the techniques that were new in this year's system, including processing of the acoustic training data, test segmentation, revised cepstral normalization and Vocal Tract Length Normalization (VTLN), band-specific models, Diagonal transform Speaker Adaptive Training (DSAT), and a modified ROVER method for system combination. We show that by combining all the above techniques, we were able to improve the recognition accuracy on the 1997 Hub-4E evaluation test by 27% relative to our 1997 system (from 20.4% to 14.8%). We also present our results on the 1998 Hub-4E and Hub-4Sp benchmarks, and discuss the differences between the English and Spanish transcription systems.

1. INTRODUCTION

The 1997 BBN BYBLOS system [1] was focused on improving the recognition accuracy of the F0 and F1 focus conditions (high fidelity prepared and spontaneous speech, respectively). Therefore, it consisted of a single set of gender-dependent acoustic models, and used a very simple segmentation procedure, based on silence detection alone. This approach gave us good performance on the F0/F1 conditions, but poor accuracy on the telephone data (F2 condition). That hampered our overall performance on the 1997 Hub-4E evaluation, since 16% of the test set consisted of telephone speech.

In the 1998 system we followed a more elaborate segmentation strategy that resulted in sizeable gains across all focus conditions, but especially on F2. Band-specific models, VTLN and DSAT adaptation gave us additional gains. In short, our effort this year was targeted to a clean and careful system design, using refined speech recognition technology from the 1997 BYBLOS system.

The paper is organized as follows. Section 2 gives an overview of the BYBLOS system used for the 1998 Hub-4E evaluation. In section 3 we discuss the improvements made to the system since the 1997 benchmark, along with experimental results. This is followed by a description of our 1998 Hub-4E evaluation results, and the computational resources used during the evaluation. We end with section 5, in which we give a brief discussion of the Spanish system.

2. SYSTEM DESCRIPTION

The monolithic broadcast news input is segmented into two band-specific episodes using a dual-band gender-independent phoneme decoder. Each channel episode is then segmented and gender-classified in one step with a dual-gender context-dependent word decoder. Speaker change detection and clustering is applied within each channel-gender sub-episode to define speaker turns and speaker clusters for unsupervised adaptation. Analysis is performed with a 36-pole LPC model, and Vocal Tract Length Normalization (VTLN). The spectrum mean and variance is normalized over each speaker turn, with speech and non-speech frames normalized separately. Speaker-Independent (SI), Gender-Dependent (GD) band-specific models are estimated from the training data, and are refined by Diagonal transform Speaker Adaptive Training (DSAT), a version of maximum likelihood SAT that uses many diagonal transformations [2, 3].

For each channel and gender we create three models: a triphone within-word Phoneme-Tied Mixture (PTM) model with 50 phonetic codebooks, 256 Gaussians per phone, and approximately 29K mixture weight vectors associated with the codebooks; a quinphone within-word State-Clustered-Tied Mixture (SCTM) model with approximately 3500 states, 64 Gaussians per state, and 30K mixture weight vectors; and a quinphone between-word SCTM model with a similar number of parameters.

SI decoding is done in two passes [4] – a forward PTM fast-match and a backward SCTM pass producing a scored word lattice that is then converted into an N-best, including trigram probabilities. The N-best is rescored with the between-word SCTM model and a trigram LM. The best hypothesis is used for unsupervised adaptation to each speaker cluster. Only the DSAT SCTM within-word model is adapted at this stage with two iterations of MLLR [5] adaptation. We also re-normalize the test spectrum for speech and non-speech based on the best hypothesis. The backward-pass decode is then repeated using the DSAT adapted model. The best hypotheses in the resulting N-best are used to adapt the DSAT between-word model. Finally, we rescore the N-best with the between-word DSAT adapted model to find the top choice.

Four decodings are done for each segment – with 125, 100, and 80 frame/second test set analysis, and a triphone decode with 100 frame/second analysis. System Combination is applied to the outputs of the four systems in order to choose the final answers.

3. RECENT IMPROVEMENTS

3.1. Speaker Change Detection

As described in [1], the segmentation procedure that we used in the 1997 Hub-4 evaluation consisted of the following simple steps:

1. Dual-gender phoneme decoding for silence and gender detection.
2. Smoothing of the recognized hypothesis to eliminate unreliable gender changes, and separation of male from female seg-

ments. Also, long silence segments are discarded.

3. Chopping of the hypothesis into small segments, averaging 4 sec. duration, using the silence detection information from step (1).
4. Clustering of the short segments into speaker clusters [6] for adaptation.
5. Cepstral mean and variance normalization applied to each segment, with speech and non-speech frames normalized separately (2-level CMS and variance normalization).

Notice that in training, the cepstral normalization is applied to each speaker turn, so the above procedure results in a mismatch between training and testing. Ideally, we would like to detect speaker changes on the test, and normalize the cepstra on speaker turns. We modified slightly the above procedure to include speaker change detection, as follows:

- After step (2) above, perform speaker change detection using the techniques described in [7].
- Apply speaker clustering to refine the speaker turns and merge short speakers for adaptation.
- Divide each speaker turn into 4-sec. segments based on the silence detection.
- Perform 2-level CMS and variance normalization on speaker turns.

This modification gave us a significant gain in recognition accuracy, as shown in Table 1, primarily because of the better normalization.

	Speaker change detection	
Condition	no	yes
F0	13.6	12.5
F1	19.4	18.9
F2	37.7	34.4
F3	28.8	30.3
F4	27.8	24.6
F5	32.3	32.7
FX	47.8	45.5
Overall	22.7	21.4

Table 1: Effect of speaker change detection and speaker turn normalization on the h4e97 test set (SI unadapted results, models trained on 80 hours of broadcast news).

3.2. Processing of Training Data

While we were processing the new 100 hours of training data for inclusion in our acoustic training, we noticed that the annotators were much more accurate in the transcription of non-speech sounds, word fragments, unclear words and overlapping speech. We decided to take advantage of the enhanced information by changing our procedure with respect to each of the above characteristics. In addition, we reprocessed the 1997 Hub-4 training data (first 100 hours) to be consistent with the new procedure.

Non-speech sounds. The only non-speech event that we used in the 1997 Hub-4 training data was laughter, denoted by the word [LAUGH]. This year we added 3 more non-speech words, namely [BREATH] for breathing, [COUGH] for coughing, and [LIPSMACK] for lipsmack sounds. The pronunciation for each of the above words is based on the specialized phonemes LGH, BRT, COU and LIP, respectively. We also introduced a new "garbage" phoneme GRB, to use in the pronunciations of unclear words and word fragments, as described below.

Word fragments. We made pronunciations for all the word fragments transcribed. For those fragments that we could clearly identify their intended word, we extracted their pronunciation from that of the intended word. For the ones that were ambiguous, we made a pronunciation consisting only of the "garbage" phoneme GRB, based on the length of the fragment.

Unclear speech. Unclear words are marked in the transcriptions in parentheses, indicating that the annotator was not sure of the spoken words. Sometimes, the unclear words are marked with an empty set of parentheses, indicating that the annotator could not identify any spoken word at all. We processed the transcriptions semi-automatically so that long segments that contained empty unclear words were divided into smaller segments, so that we could discard the segments with the empty unclear words without significant loss in training data. For each non-empty unclear word, we made a pronunciation using again the "garbage" phoneme GRB based on the length of the unclear word.

Overlapping speech. In the new 100 hours of training data, overlapping speech is clearly marked, so we can easily discard the segments on the overlapping region. However, the 1997 Hub-4 training data does not have this markers, so we had to remove the overlapping regions by using a semi-automatic method.

Speaker turn boundary adjustment. On many occasions we have found that the transcribers of the Hub-4 broadcast news training data mark the beginning or end of a speaker turn to be very close to the beginning or end of speech, respectively, without leaving any noticeable silence. This causes a problem in our system, which expects to find silence frames at the endpoints of each training segment. Also, in some cases the boundary between two speaker turns is not correctly marked, resulting in misaligned training segments that corrupt our acoustic models.

In an attempt to correct most of the annotation errors of the type described above, we developed an automatic method for adjusting the marked boundaries between speaker turns. The method is based on the assumption that there is always a number of silence frames between two consecutive speaker turns; if we could detect the location of this silence, we could re-position the boundary to be at the middle of the silence, dividing the silence frames equally between successive speakers. Given the transcription for a speaker turn, we can run a constrained decode to obtain the most likely word begin/end times, according to the acoustic model in use. Therefore, we can apply the same procedure to reposition the boundary between two speaker turns, as follows:

- For any two consecutive turns A and B, create a "merged" turn C whose begin time is the begin time of A, and its end time is the end time of B. The transcription for turn C is the concatenation of the A and B transcriptions.
- Rescore each merged turn to obtain word time marks and the position of silences. Obviously, the acoustic model used for

rescoring needs to be gender-independent (GI), since one male speaker may be followed by a female, and vice-versa.

- If we detect a silence between the last word of turn A and the first word of turn B, then we split the merged turn C at the middle of the silence, adjusting the corresponding boundary.
- If no silence can be detected between the two turns, we use the word time-stamps to adjust the boundary.

We were surprised to find that among all the above processing techniques, only the segment boundary adjustment and the modeling of breath gave us a gain, as shown in Table 2.

Type of processing	Accuracy
1997 Baseline	21.4
1998 trans. processing	21.3
+ boundary adjustment	20.9
+ breath modeling	20.5

Table 2: Effect of training transcription processing on the h4e97 test set (SI unadapted results, models trained on 80 hours of broadcast news).

3.3. Inclusion of More Training Data

By adding the new 100 hours of data into our acoustic training, processed as described in the previous section, we achieved a 0.9% absolute gain over the 80 hours baseline. Notice that because of commercials, bad data, etc. the total actual amount of training we used for the 1998 Hub-4 evaluation was 150 hours. Table 3 shows the effect of the new training data on each condition of the h4e97 test set.

	Amount of training	
Condition	80 hours	150 hours
F0	11.8	11.3
F1	17.8	17.4
F2	34.1	32.7
F3	29.1	27.2
F4	23.9	23.0
F5	28.1	25.1
FX	44.3	41.0
Overall	20.5	19.6

Table 3: Effect of the new 100 hours (70 actually used) of training on the h4e97 test set (SI unadapted results).

3.4. Test Segmentation

The speaker change detection algorithm that we discussed previously fails to detect a speaker change when there is no sufficiently long silence at the boundary. This results often in segments that contain mixed bands and/or genders. The solution is to detect channel changes first, and then detect gender changes within each channel, independently. In the 1998 system we used a dual-band gender independent, context independent phone-class model for band detection, trained using the following procedure:

- first, we label all the training data with a regular gender dependent (GD) quinphone SCTM between-word model, to obtain a phoneme and state alignment for each frame.
- phonemes are clustered into 8 broad classes: voiced continuents, fricatives and sibilants, obstruents, breath, laughter, lipsmack, music, and silence. The labels are modified so that they map each frame into one of the above phone classes.
- then, we separate the wideband from the narrow-band segments. We determine which segments are narrow-band by computing the ratio of the average energy of the segment that is above 4 kHz to the energy between 125 Hz and 4 kHz, and comparing that ratio with a fixed threshold.
- the labels for the narrow band segments are modified again so that the first five phone classes in the list above are mapped to a different name, to indicate clearly that these are narrow-band phones. This results in a set of labels with a total of 13 distinct phone classes.
- finally, we train a gender independent (GI), context independent phone class model, using the above labels. During the training, we constrain the forward-backward procedure to keep the state alignment fixed, according to the labels.

The dual-band phone class model is very good in detecting band changes, but the silence and music detection is not very accurate, most likely because of the overlap between the small number of phone classes. We can trust the detected long silences to chop into relatively long segments (30 sec. on the average), but not shorter than that. If we use a word decoder to detect silences, we can alleviate the problem of recognizing silence in the middle of a word. Therefore, we decided that we should use a word decoder with a small vocabulary (5000 words) to detect silences after separating the channels.

Our system is gender dependent, so in addition to the silence detection, we need to perform a gender detection as well. Therefore, we decided to train a dual-gender triphone PTM within-word model and a dual-gender quinphone SCTM within-word model, to be used in the forward and backward passes of the word decoder. The training of the dual-gender models is performed as follows:

- the standard set of phonemes is duplicated (except for the silence phoneme) and for each non-silence phoneme we have a male and a female version.
- the training dictionary is modified so that each word has two versions, male and female, and each version has a pronunciation that consists of the corresponding GD phonemes.
- the training labels are modified so that each phoneme is mapped to its corresponding GD version.
- the training transcriptions are modified so that each word is mapped to its corresponding GD version.
- everything else in the training procedure is the same as in the standard case.

By combining the above two models, we defined the following segmentation strategy:

1. Phone-class decode for band and long silence detection

2. Smooth the recognized hypothesis from step (1), to eliminate false alarms (unreliable band changes), and separate the wide-band segments from the narrow-band ones.
3. Perform gender detection and more accurate silence detection on each band separately, using dual-gender word decoding. Before recognition, the channel turns are divided into segments averaging 30 sec. duration, using the silence detection information from step (1). For each such segment, we obtain a hypothesis consisting of male and/or female words. Smoothing is performed on each channel turn to eliminate unreliable gender changes. Also, long silence segments are discarded.
4. Speaker change detection is applied on each channel-gender turn.
5. After the speaker changes have been determined, we perform speaker clustering to reduce the number of speakers and merge short speaker turns for adaptation.
6. Divide each speaker turn into 4-sec. segments based on the silence detection from step (3).
7. Perform 2-level CMS and variance normalization on speaker turns.

Table 4 shows that there is a 0.8% absolute gain from using the above segmentation procedure. We can see that most of the gain comes from the improvement on the F2 condition. This indicates that the band detection using the dual-band phone class decoder does a good job in separating wideband from narrow-band speech.

	Band detection	
Condition	no	yes
F0	11.3	11.5
F1	17.4	17.5
F2	32.7	28.3
F3	27.2	26.2
F4	23.0	23.0
F5	25.1	24.1
FX	41.0	39.3
Overall	19.6	18.8

Table 4: Effect of band detection and more accurate silence detection on the h4e97 test set (SI unadapted results, models trained on 150 hours of broadcast news).

3.5. Language Modeling

All the results shown so far have been using a 44k recognition lexicon, and a language model identical to the one used in the 1997 system. We found that there is a small gain for adding acronyms into the recognition lexicon as compound words, and using a 60k lexicon as well. So at this point, we decided to retrain the acoustic models with the compound acronyms, and also add more training data to the language model.

We used a total of 600 million words to train the language model. The data are from the following sources:

- 556 million words selected from the LDC official releases "North American News Text Corpus", "North Amerian News Text Corpus (Supplement) and AP Worldstream English", and the previous release in 1997. We excluded data prior to 1994. Of course, we excluded data in the test epochs (1996/10/15 to 1996/11/15, and after 1998/02/28).
- 40 million words from in-house data obtained through Primary Source Media
- 4 million words from Hub-4 acoustic training data.

The resulting LM had 13M bigrams and 43M trigrams. The transcriptions for the acoustic data were included 20 times in the LM training. Coverage of the 60k lexicon on the H4E97 test set was 99.41%.

As shown in Table 5, the overall gain from retraining the acoustic models with compound acronyms, increased LM training, and 60k recognition lexicon was 0.5%.

Condition	44k LM	60k LM + acronyms
F0	11.5	11.2
F1	17.5	17.0
F2	28.3	29.1
F3	26.2	24.6
F4	23.0	22.0
F5	24.1	22.4
FX	39.3	36.5
Overall	18.8	18.3

Table 5: Effect of larger recognition lexicon and LM on the h4e97 test set (SI unadapted results, models trained on 150 hours of broadcast news).

3.6. Band Specific Models and Vocal Tract Length Normalization

In a previous section we showed that by performing band detection during the segmentation stage, we were able to reduce the word error rate on the telephone data, mainly because of the better cepstral normalization. It is well known that additional gain can be achieved by using band-specific models for recognition. Therefore, we trained a second (narrow-band) set of gender dependent models by reducing the bandwidth data analysis to the range of the telephone data. In Table 6 we can see that the use of band-specific models gave us a 5% absolute gain on F2, and an overall gain of 0.9%.

In addition, we tried VTLN on the test only, by estimating a stretch factor for each speaker cluster, using a Gaussian mixture model (GMM) against which speakers were scored at a multiplicity of warps. To compensate for the inherent GMM likelihood bias for cepstra at different warps, the determinant of the VTL transformation was estimated empirically per speaker and applied [3]. We observed only a 0.4% gain from VTLN on the test, as shown in Table 6. We tried to apply VTLN in training as well, but it didn't give us any additional gain. Due to time constraints, we decided to use VTLN on the test only.

3.7. Adaptation

As we mentioned in section 2, the 1998 system used DSAT models for adaptation. For each speaker cluster, we estimated a set of 8 full-matrix MLLR transformations, based on the hypotheses from the SI unadapted recognition pass. Then, for each speaker we applied the

Condition	Band-specific models		
	W only	W + N	W + N + VTLN
F0	11.2	11.0	10.8
F1	17.0	17.0	17.2
F2	29.1	24.1	23.2
F3	24.6	24.1	23.3
F4	22.0	23.3	22.8
F5	22.4	23.3	21.0
FX	36.5	35.3	33.7
Overall	18.3	17.4	17.0

Table 6: Effect of band-specific models (W = wide, N = narrow) and VTLN on the h4e97 test set (SI unadapted results, models trained on 150 hours of broadcast news).

estimated transformation to the DSAT SCTM within-word models and ran the backward pass. The best hypotheses from the backward adapted decoding were used to adapt the DSAT SCTM between-word models for the final N-best rescoring pass. This adaptation procedure gave us a 9.4% relative overall gain, as shown in Table 7.

We found that we could gain an additional 0.2% by re-normalizing the cepstra before adaptation, based on the newly detected silence positions from the SI unadapted recognition pass. In light of this result, we decided to go one step further and re-segment each speaker turn based on the newly detected silences, under the hypothesis that the silence detection from the SI unadapted decoding is better than the one from the dual-gender word decoding performed during the segmentation pass. Although re-segmenting the speaker turns gave a gain on the F2 and FX conditions, the overall effect was slightly worse, so we didn't use this approach in the final system.

Condition	SI	DSAT Adapted		
		standard	re-norm	re-seg + re-norm
F0	10.8	9.9	9.8	10.2
F1	17.2	15.1	14.8	15.4
F2	23.2	19.9	19.3	18.8
F3	23.3	22.3	22.2	23.3
F4	22.8	21.4	21.1	20.8
F5	21.0	19.7	18.9	20.5
FX	33.7	31.9	31.7	30.3
Overall	17.0	15.4	15.2	15.3

Table 7: Effect of DSAT MLLR adaptation and cepstral re-normalization on the h4e97 test set (models trained on 150 hours of broadcast news).

3.8. System Combination

We applied a modified ROVER [8] algorithm, that uses a weighted selection procedure which is optimized to reduce word error rate [9]. The above technique was applied to the outputs of four systems:

- The quinphone system with 80 frames/sec. test set analysis
- The quinphone system with 100 frames/sec. test set analysis
- The quinphone system with 120 frames/sec. test set analysis
- The triphone system with 100 frames/sec. test set analysis

Table 8 shows that by combining the above systems we could achieve a 0.4% reduction in WER.

Condition	Baseline	System Combination
F0	9.8	9.6
F1	14.8	14.8
F2	19.3	18.6
F3	22.2	22.4
F4	21.1	21.0
F5	18.9	18.4
FX	31.7	29.5
Overall	15.2	14.8

Table 8: Effect of system combination on the h4e97 test set (DSAT adapted results, models trained on 150 hours of broadcast news).

4. 1998 HUB-4E RESULTS

Table 9 shows the BBN results on the 1998 Hub-4E benchmark. The benchmark consisted of two sets; set 1 was recorded in the same epoch as the H4E97 test set, while set 2 was recorded in a later period. We can see that the performance on set 1 is comparable to the one achieved on the H4E97 development test.

Test	F0	F1	F2	F3	F4	F5	FX	all
set1	7.4	15.6	23.2	24.1	10.5	17.6	23.4	15.2
set2	9.5	13.9	18.4	16.2	17.0	18.6	38.2	14.2

Table 9: BBN results in the 1998 Hub-4 primary evaluation benchmark

It is also interesting to see the overall improvement of the BBN BYBLOS system in 1998. Using the techniques described in the previous section, we were able to reduce the word error rate on the H4E97 test set by 27% relative to our 1997 system, demonstrated in Table 10 by focus condition.

Condition	1997 system	1998 system	Rel. gain (%)
F0	12.3	9.6	22
F1	17.8	14.8	17
F2	32.6	18.6	43
F3	27.9	22.4	20
F4	24.7	21.0	15
F5	28.2	18.4	35
FX	42.8	29.5	31
Overall	20.4	14.8	27

Table 10: Comparison of the BYBLOS 1997 and 1998 systems on h4e97.

Computational Resources The computation for this evaluation was done on Intel-based PCs with 450MHz Pentium-II CPUs, 512MB of RAM, and 1.2GB of swap space. The operating system was Linux 2.0.3x and the compiler was GNU gcc version egcs-2.91.6 from the Free Software Foundation. These machines have a

1995 SPEC base rating of 17.2 for the integer test, and 11.8 for the floating point test.

Table 11 shows timing information for the basic recognition stages of the 1998 primary system on the H4E98 test sets.

Stage	xRT
Segmentation	10
Unadapted decoding (4 systems)	86
Adaptation (4 systems)	44
Adapted decoding (4 systems)	103
System Combination	1
Total	244

Table 11: Timing information of the 1998 BYBLOS system on the H4E98 test sets (measured on Pentium-II 450MHz PC's with 512MB RAM, 1.2GB swap, running Linux 2.0.3x OS).

5. SPANISH BROADCAST NEWS

For the 1998 Hub-4Sp benchmarck, we tried to apply the technology and procedures developed for the English system. However, due to the different amount of acoustic and language training data between the Spanish and English Systems, as well as the limited time frame imposed by the evaluation, we had to simplify the Spanish system slightly.

In particular, we did not train narrow-band models, neither used VTLN nor system combination; the wideband quinphone SCTM models had fewer states, because of the small amount of acoustic training data available; and the language model had much fewer estimated bigram and trigram probabilities. These differences are shown in detail in Table 12.

System Feature	English	Spanish
Acoustic Training	150 hours	26 hours
Language Training	600M words	157M words
Quinphone SCTM states	3500	2000
LM bigrams	13M	3.0M
LM trigrams	43M	5.5M
H4E97 coverage	99.4%	98.4%
Narrow-band models	Yes	No
VTLN	Yes	No
System Combination	Yes	No

Table 12: Differences between the 1998 English and Spanish systems

Despite the above differences, the BBN Hub-4 Spanish system attained the best performance on the 1998 Hub-4Sp test set, and the result is shown in Table 13.

Notice that there was no presence of F2 or F5 data in the test set. The huge word error rate on FX is because there are only 6 words in the FX condition, but our system inserted 28 words over several noisy regions of non-speech events.

Condition	WER
F0	14.0
F1	40.1
F2	N/A
F3	25.4
F4	19.8
F5	N/A
FX	566.7
Overall	21.5

Table 13: BBN results in the 1998 Hub-4 Spanish evaluation benchmark

6. ACKNOWLEDGMENTS

This work was supported by the Defense Advanced Research Projects Agency and monitored by Ft. Huachuca under contract No. DABT63-94-C-0063 and by the Defense Advanced Research Projects Agency and monitored by NRaD under contract No. N66001-97-D-8501. The views and findings contained in this material are those of the authors and do not necessarily reflect the position or policy of the Government and no official endorsement should be inferred.

References

1. F. Kubala, J. Davenport, et al., "The 1997 BBN BYBLOS System applied to Broadcast News Transcription", *Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop*, Lansdowne VA, Feb. 1998.
2. T. Anastasakos, J. McDonough, R. Schwartz, and J. Makhoul, "A Compact Model for Speaker-Adaptive Training", *ICSLP Proceedings*, Philadelphia, PA., October 3-6 1996.
3. G. Zavaliagkos, J. Billa, et al., "The BBN/Byblos Hub-5 System Description", *NIST 1998 Hub-5 Workshop*, Linthicum, Maryland, Sep. 1998.
4. L. Nguyen and R. Schwartz, "Efficient 2-Pass N-Best Decoder", *Proc. EuroSpeech '97*, Rhodes, Greece, Sep. 1997, pp. 167-170.
5. C. J. Leggetter, P. C. Woodland, "Flexible Speaker Adaptation Using Maximum Likelihood Linear Regression", *Proceedings of the Spoken Language Systems Technology Workshop*, Austin TX, Jan. 1995, pp. 110-115.
6. H. Jin, F. Kubala, R. Schwartz, "Automatic Speaker Clustering", *Proceedings of the DARPA Speech Recognition Workshop*, Chantilly VA, Feb. 1997.
7. D. Liu, F. Kubala, "Fast Speaker Change Detection for Broadcast News Transcription and Indexing", submitted for publication to Eurospeech '99, Budapest, Hungary, Sep. 99.
8. J. G. Fiscus, "A Post-Processing System to Yield Reduced Word Error Rates: Recognizer Output Voting Error Reduction (ROVER)", *Draft. 1997 LVCSR DARPA HUB-5E Workshop*, Linthicum, Maryland, May 13-15, 1997.
9. M. Siu, H. Gish, F. Richardson, "Improved Estimation, Evaluation and Applications of Confidence Measures for Speech Recognition", *Proc. EuroSpeech '97*, Rhodes, Greece, Sep. 1997.