# Generating Training Data for Medical Dictations

Sergey Pakhomov
University of Minnesota, MN
pakhomov.sergey@mayo.edu

Michael Schonwetter
Linguistech Consortium, NJ
MSchonwetter@qwest.net

Joan Bachenko
Linguistech Consortium,NJ
bachenko@mnic.net

## Abstract

In automatic speech recognition (ASR) enabled applications for medical dictations, corpora of literal transcriptions of speech are critical for training both speaker independent and speaker adapted acoustic models. Obtaining these transcriptions is both costly and time consuming. Non-literal transcriptions, on the other hand, are easy to obtain because they are generated in the normal course of a medical transcription operation. This paper presents a method of automatically generating texts that can take the place of literal transcriptions for training acoustic and language models. ATRS[1] is an automatic transcription reconstruction system that can produce near-literal transcriptions with almost no human labor. We will show that (i) adapted acoustic models trained on ATRS data perform as well as or better than adapted acoustic models trained on literal transcriptions (as measured by recognition accuracy) and (ii) language models trained on ATRS data have lower perplexity than language models trained on non-literal data.

## Introduction

Dictation applications of automatic speech recognition (ASR) require literal transcriptions of speech in order to train both speaker independent and speaker adapted acoustic models. Literal transcriptions may also be used to train stochastic language models that need to perform well on spontaneous or disfluent speech. With the exception of personal desktop systems, however, obtaining these transcriptions is costly and time consuming since they must be produced manually by humans educated for the task. The high cost makes literal transcription unworkable for ASR applications that require adapted acoustic models for thousands of talkers as well as accurate language models for idiosyncratic natural speech.

Non-literal transcriptions, on the other hand, are easy to obtain because they are generated in the normal course of a medical transcription operation. It has been previously shown by Wightman and Harder (1999) that the non-literal transcriptions can be successfully used in acoustic adaptation. However, non-literal transcriptions are incomplete. They exclude many utterances that commonly occur in medical dictation—filled pauses, repetitions, repairs, ungrammatical phrases, pleasantries, asides to the transcriptionist, etc. Depending on the talker, such material may constitute a significant portion of the dictation.

We present a method of automatically generating texts that can take the place of literal transcriptions for training acoustic and language models. ATRS is an automatic transcription reconstruction system that can produce near-literal transcriptions with almost no human labor.

The following sections will describe ATRS and present experimental results from language and acoustic modeling. We will show that (i) adapted acoustic models trained on ATRS data perform as well as or better than adapted acoustic models trained on literal transcriptions (as measured by recognition accuracy) and (ii) language models trained on ATRS data have lower perplexity than language models trained on non-literal data. Data used in the experiments comes from medical dictations. All of the dictations are telephone speech.

[1] patent pending (Serial No.: 09/487398)

## 1 Dictation Applications of ASR

The application for our work is medical dictation over the telephone. Medical dictation differs from other telephony based ASR applications, e.g. airline reservation systems, because the talkers are repeat users and utterances are long. Dictations usually consist of 1-30 minutes of speech. The talkers call in 3-5 days per week and produce between 1 and 12 dictations each day they call. Hence a medical dictation operation has access to hours of speech for each talker.

Spontaneous telephone speech presents additional challenges that are caused partly by a poor acoustic signal and partly by the disfluent nature of spontaneous speech. A number of researchers have noted the effects of disfluencies on speech recognition and have suggested various approaches to dealing with them at language modeling and post-processing stages. (Shriberg 1994, Shriberg 1996, Stolcke and Shriberg 1996, Stolcke et al. 1998, Shriberg and Stolcke 1996, Siu and Ostendorf 1996, Heeman et al. 1996) Medical over-the-telephone dictations can be classified as spontaneous or quasi-spontaneous discourse (Pakhomov 1999, Pakhomov and Savova 1999). Most physicians do not read a script prepared in advance, instead, they engage in spontaneous monologues that display the full spectrum of disfluencies found in conversational dialogs in addition to other "disfluencies" characteristic of dictated speech. An example of the latter is when a physician gives instructions to the transcriptionist to modify something in the preceding discourse, sometimes as far as several paragraphs back.

Most ASR dictation applications focus on desktop users; for example, Dragon, IBM, Philips and Lernout & Hauspie all sell desktop dictation recognizers that work on high quality microphone speech. Typically, the desktop system builds an adapted acoustic model if the talker "enrolls", i.e. reads a prepared script that serves as a literal transcription. Forced alignment of the script and the speech provides the input to acoustic model adaptation.

Enrollment makes it relatively easy to obtain literal transcriptions for adaptation. However, enrollment is not feasible for dictation over the telephone primarily because most physicians will refuse to take the time to enroll. The alternative is to hire humans who will type literal transcriptions of dictation until enough have been accumulated to build an adapted model, an impractical solution for a large scale operation that processes speech from thousands of talkers. ATRS is appealing because it can generate an approximation of literal transcription that can replace enrollment scripts and the need for manually generated literal transcriptions.

## 2 Three Classes of Training Data

In this paper, training texts for language and acoustic models fall into three categories:

**Non-Literal:** Non-literal transcripts present the meaning of what was spoken in a written form appropriate for the domain. In a commercial medical transcription operation, the non-literal transcript will present the dictation in a format appropriate for a medical record. This typically involves (i.) ignoring filled pauses, pleasantries, and repeats; (ii.) acting on directions for repairs ("delete the second paragraph and put this in instead..."); (iii.) adding non-dictated punctuation; (iv.) correcting grammatical errors; and (v.) re-formatting certain phrases such as "Lung are Clear", to a standard form such as "Lungs - Clear".

**Literal:** Literal transcriptions are exact transcriptions of what was spoken. This includes any elements not found in the non-literal transcript, such as filled pauses (um's and ah's), pleasantries and body noises ("thank you very much, just a moment, cough"), repeats, fragments, repairs and directions for repairs, and asides ("make that bold"). Literal transcriptions require significant human effort, and therefore are expensive to produce. Even though they are carefully prepared, some errors will be present in the result.

In their study of how humans deal with transcribing spoken discourse, Lindsay and O'Connell (1995) have found that literal transcripts were "far from verbatim." (p.111) They find that the transcribers in their study tended to have the most difficulty transcribing hesitation phenomena, followed by sentence fragments, adverbs and conjunctions and, finally, nouns, verbs, adjectives and prepositions.

Our informal observations made from the transcripts produced by highly trained medical transcriptionists suggest approximately 5% error margin and a gradation of errors similar to the one found by Lindsay and O'Connell.

**Semi-Literal:** Semi-literal transcripts are derived using non-literal transcripts, the recognizer output, a set of grammars, a dictionary, and an interpreter to integrate the recognized material into the non-literal transcription. Semi-literal transcripts will more closely resemble the literal transcripts, as many of the elements missing from the non-literal transcripts will be restored.

## 3 Model Adaptation

It is well known that ASR systems perform best when acoustic models are adapted to a particular talker's speech. This is why commercial desktop systems use enrollment. Although less widely applied, language model adaptation based on linear interpolation is an effective technique for tailoring stochastic grammars to particular domains of discourse and to particular speakers (Savova et al. (2000), Weng et al. (1997)).

The training texts used in acoustic modeling come from recognizer-generated texts, literal transcriptions or non-literal transcriptions. Within the family of transformation and combined approaches to acoustic modeling (Digalakis and Neumeyer (1996), Strom (1996), Wightman and Harder (1999), Hazen and Glass (1997)) three basic adaptation methods can be identified: unsupervised, supervised, or semi-supervised. Each adaptation method depends on a different type of training text. What follows will briefly introduce the three methods.

**Unsupervised adaptation** relies on the recognizer's output as the text guiding the adaptation. Efficacy of unsupervised adaptation fully depends on the recognition accuracy. As Wightman and Harder (1999) pointed out, unsupervised adaptation works well in laboratory conditions when the speech signal has large bandwidth and is relatively "clean" of background noise, throat clearings, and other disturbances. In laboratory conditions, the errors introduced by unsupervised adaptation can be averaged out by using more data (Zavaliagkos and Colthurst, 1997); however, in a telephony operation with degraded input that is not feasible.

**Supervised adaptation** is dependent on literal transcription availability and is widely used in enrollment in most desktop ASR systems. A speaker's speech sample is transcribed verbatim and then the speech signal is aligned with pronunciations frame by frame for each individual word. A speaker independent model is augmented to include the observations resulting from the alignment.

**Semi-supervised adaptation** rests on the idea that the speech signal can be partially aligned by using of the recognition output and the non-literal transcription. A significant problem with semi-supervised adaptation is that only the speech that the recognizer already recognizes successfully ends up being used for adaptation. This reinforces what is already well represented in the model. Wightman and Harder (1999) report that semi-supervised adaptation has a positive side effect of excluding those segments of speech that were mis-recognized for reasons other than a poor acoustic model. They note that background noise and speech disfluency are detrimental to the unsupervised adaptation.

In addition to the two problems with semi-supervised adaptation pointed out by Wightman and Harder, we find one more potential problem. As a result of matching the word labels produced by the recognizer and the non-literal transcription, some words may be skipped which may introduce unnatural phone transitions at word boundaries.

**Language model adaptation** is not an appropriate domain for acoustic adaptation methods. However, adapted language models can be loosely described as supervised or unsupervised, based on the types of training texts—literal or non-literal—that were used in building the model.

In the following sections we will describe the system of generating data that is well suited for acoustic and language adaptation and present results of experimental evaluation of this system.

## 3.2 Generating semi-literal data

ATRS is based on reconstruction of non-literal transcriptions to train utterance specific language models. First, a non-literal transcription is used to train an augmented probabilistic finite state model (APFSM) which is, in turn, used by the recognizer to re-recognize the exact same utterance that the non-literal transcription was generated from. The APFSM is constructed by linear interpolation of a finite state model where all transitional probabilities are equal to 1 with two other stochastic models.

One of the two models is a background model that accounts for expressions such as greetings, thanking, false starts and repairs. A list of these out-of-transcription expressions is derived by comparing already existing literal transcriptions with their non-literal transcription counterparts. The other model represents the same non-literal transcription populated with filled pauses (FP) ("um's and ah's") using a stochastic FP model derived from a relatively large corpus of literal transcriptions (Pakhomov, 1999, Pakhomov and Savova, 1999).

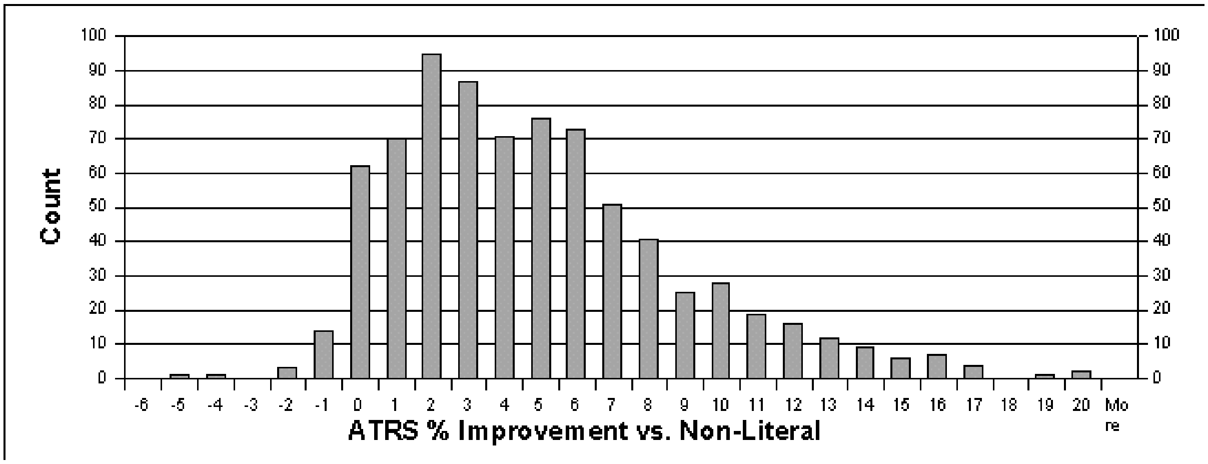

**Figure { SEQ Figure * ARABIC } Percent improvement in true data representation of ATRS reconstruction vs. Non-Literal data**

Interpolation weights are established empirically by calculating the resulting model's perplexity against held out data. Out-of-vocabulary (OOV) items are handled provisionally by generating on-the-fly pronunciations based on the existing dictionary spelling-pronunciation alignments. The result of interpolating these two background models is that some of the transitional probabilities found in the finite state model are no longer 1.

The language model so derived can now be used to produce a transcription that is likely to be more true to what has actually been said than the non-literal transcription that we started to work with.

Further refinement of the new semi-literal transcription is carried out by using dynamic programming alignment on the recognizer's hypothesis (HYP) and the non-literal transcription that is used as reference (REF). The alignment results in each HYP label being designated as a MATCH, a DELETION, a SUBSTITUTION or an INSERTION. Those labels present in the HYP stream that do not align with anything in the REF stream are designated as insertions and are assumed to represent the out-of-transcription elements of the dictation. Those labels that do align but do not match are designated as substitutions. Finally, the labels found in the REF stream that do not align with anything in the HYP stream are designated as deletions.

The final semi-literal transcription is constructed differently depending on the intended purpose of the transcription. If the transcription will be used for acoustic modeling, then the MATCHES, the REF portion of SUBSTITUTIONS and the HYP portion of only those INSERTIONS that represent punctuation and filled pauses make it into the final semi-literal transcription. It is important to filter out everything else because acoustic modeling is very sensitive to misalignment errors. Language modeling, on the other hand, is less sensitive to alignment errors; therefore, INSERTIONS and DELETIONS can be introduced into the semi-literal transcription.

One method of ascertaining the quality of semi-literal reconstruction is to measure its alignment errors against literal data using a dynamic programming application. By measuring the correctness spread between ATRS and literal data, as well as the correctness spread between non-literal and literal data, the ATRS alignment correctness rate was observed to be 4.4% higher absolute over 774 dictation files tested. Chart 1 summarizes the results. The X axis represents the number of dictations in each bin displayed along the Y axis representing the % improvement over the non-literal counterparts. The results showed nearly all ATRS files had better alignment correctness than their non-literal counterparts. The majority of the reconstructed dictations resemble literal transcriptions between 1% and 8% better than their non-literal counterparts. These results are statistically significant as evidenced by a t-test at 0.05 confidence level. Much of the increase in alignment can be attributed to the introduction of filled pauses by ATRS. However, ignoring filled pauses, we have observed informally that the correctness still improves in ATRS files versus non-literal.

In the following sections we will address acoustic and language modeling and show that semi-literal training data is a good substitute for literal data.

## 4 Experimental results

The usefulness of semi-literal transcriptions was evaluated in two ways: acoustic adaptation and language modeling.

### 4.1 Adapted acoustic model evaluation

Three speaker adapted acoustic models were trained for each of the 5 talkers in this study using the three types of label files and evaluated on the talker's testing data.

### 4.1.1 Setup

The data collected for each talker were split into testing and training.

#### *Training Data*

45-55 minutes of audio data was collected for each of the six talkers in this experiment:

| | |
|---|---|
| A | female |
| B | female |
| C | male |
| D | male |
| F | female |

All talkers are native speakers of English, two males and three females.

**Non-literal transcriptions** of this data were obtained in the course of normal transcription operation where trained medical transcriptionists record the dictations while filtering out disfluency, asides and ungrammatical utterances.

**Literal transcriptions** were obtained by having 5 medical transcriptionists specially trained not to filter out disfluency and asides transcribe all the dictations used in this study.

**Semi-literal transcriptions** were obtained with the system described in section 5 of this paper.

#### *Testing Data*

Three dictations (0.5 – 2 min) each were pulled out of the Literal transcriptions training set and set aside for each talker for testing.

#### *Recognition and evaluation software and formalism*

Software licensed from Entropic Laboratory was used for performing recognition, evaluating accuracy and acoustic adaptation. (Valtchev, et al. (1998)). Adapted models were trained using MLLR technique (Legetter and Woodland, (1996)) available as part of the Entropic package.

Recognition accuracy and correctness reported in this study were calculated according to the following formulas:

(1) Acc = hits – insertions / total words
(2) Correctness = hits / total words

### 4.1.2 Experiment

The following Acoustic Models were trained via adaptation with a general SI model for each talker using all available data (except for the testing data). Each model's name reflects the kind of label data that was used for training.

LITERAL

Each audio file was aligned with the corresponding literal transcription.

NON-LITERAL

Each audio file was recognized using SI acoustic and language models. The recognition output was aligned with the non-literal transcription using dynamic programming. Only those portions of audio that corresponded to direct matches in the alignment were used to produce alignments for acoustic modeling. This method was originally used for medical dictations by Wightman and Harder (1999).

SEMI-LITERAL

Each audio file has been processed to produce a semi-literal transcription that was then aligned with recognition output generated in the process of creating semi-literal transcriptions. The portions of the audio corresponding to matching segments were used for acoustic adaptation training.

The SI model had been trained on all available at the time (12 hours)[2] similar medical dictations to the ones used in this study. The data for the speakers in this study were not used in training the SI model.

### 4.1.3 Results

Table 1 shows the test results. As expected, both recognition accuracy and correctness increase with any of the three kinds of adaptation. Adaptation using Literal transcriptions yields an overall 10.84% absolute gain in correctness and 11.49% in accuracy over the baseline.

Adaptation using Non-literal transcriptions yields an overall 6.36 % absolute gain in correctness and 5.23 % in accuracy over the baseline. Adaptation with Semi-literal transcriptions yields an overall 11.39 % absolute gain in correctness and 11.05 % in accuracy over the baseline. No statistical significance tests were performed on this data.

| | **Baseline (SI) %** | | **Literal %** | | **Semi-literal %** | | **Non-literal %** | |
|---|---|---|---|---|---|---|---|---|
| **Talker** | Cor | Acc | Cor | Acc | Cor | Acc | Cor | Acc |
| **A** | 58.76 | 48.47 | 66.57 | 58.09 | 68 | 58.28 | 64.76 | 51.8 |
| **B** | 41.28 | 32.2 | 58.36 | 49.46 | 64.59 | 56.22 | 55.87 | 44.66 |
| **C** | 57.22 | 54.99 | 64.38 | 61.54 | 61.25 | 59.31 | 60.65 | 58.71 |
| **D** | 56.86 | 51.47 | 68.69 | 63.3 | 65.91 | 59.13 | 64.69 | 58.26 |
| **F** | 54.83 | 43.69 | 61.97 | 53.57 | 64.7 | 54.41 | 61.13 | 48.73 |
| | | | | | | | | |
| **AVG** | 52.49 | 44.81 | 63.33 | 56.3 | 63.81 | 55.86 | 58.85 | 50.04 |

**Table 1. Recognition results for three adaptation methods**

### 4.1.4 Discussion

The results of this experiment provide additional support for using automatically generated semi-literal transcriptions as a viable (and possibly superior) substitute for literal data. The fact that three SEMI-LITERAL adapted AM's out of 5 performed better than their LITERAL counterparts seems to indicate that there may be undesirable noise either in the literal transcriptions or in the corresponding audio. It may also be due to the relatively small amount of training data used for SI modeling thus providing a baseline that can be improved with little effort. However, the results still indicate that generating semi-literal transcriptions may help eliminate the undesirable noise and, at the same time, get the benefits of broader coverage that semi-literal transcripts can afford over NON-LITERAL transcriptions.

[2] Although 50-100 hours of data for SI modeling is the industry standard, the population we are dealing with is highly homogeneous and reasonable results can be obtained with lesser amount of data.

## 4.2 Language Model Evaluation

For ASR applications where there are significant discrepancies between an utterance and its formal transcription, the inclusion of literal data in the language model can reduce language model perplexity and improve recognition accuracy. In medical transcription, the non-literal texts typically depart from what has actually been said. Hence if the talker says "lungs are clear" or "lungs sound pretty clear", the typed transcription is likely to have "Lungs - clear". In addition, as we noted earlier, the non-literal transcription will omit disfluencies and asides and will correct grammatical errors.

Literal and semi-literal texts can be added onto language model training data or interpolated into an existing language model. Below we will present results of a language modeling experiment that compares language models built from literal, semi-literal and non-literal versions of the same training set. The results substantiate our claim that automatically generated semi-literal transcription can lead to a significant improvement in language model quality.

In order to test the proposed method's suitability for language modeling, we constructed three trigram language models and used perplexity as the measure of the models' goodness.

### Setup

The following models were trained on three versions of a 270,000-word corpus. The size of the training corpus is dictated by availability of literal transcriptions. The vocabulary was derived from a combination of all three corpora to keep the OOV rate constant.

LLM – language model built from a corpus of literal transcriptions
NLM – language model built from non-literal transcriptions
SLM – language model built from semi-literal transcriptions

Approximately 5,000-word literal transcriptions corpus consisting of 24 dictations was set aside for testing

### Results

The results of perplexity tests of the three models on the held-out data at 3-gram level are summarized in Table 2. The tests were carried out using the Entropic Transcriber Toolkit

It is apparent that SLM yields considerably better perplexity than NLM, which indicates that although semi-literal transcriptions are not as good as actual literal transcriptions, they are more suitable for language modeling than non-literal transcriptions. These results are obtained with 270,000 words of training data; however, the typical amount is dozens of million. We would expect the differences in perplexity to become smaller with larger amounts of training data.

| | Perplexity | OOV rate (%) |
|---|---|---|
| LLM | 185 | 2.61 |
| NLM | 613 | 2.61 |
| SLM | 313 | 2.61 |

**Table 2. Perplexity tests on LLM, NLM, SLM**

## Conclusions and future work

We have described ATRS, a system for reconstructing semi-literal transcriptions automatically. ATRS texts can be used as a substitute for literal transcriptions when the cost and time required for generating literal transcriptions are infeasible, e.g. in a telephony based transcription operation that processes thousands of acoustic and language models. Texts produced with ATRS were used in training speaker adapted acoustic models, speaker independent acoustic models and language models. Experimental results show that models built from ATRS training data yield performance results that are equivalent to those obtained with models trained on literal transcriptions. In the future, we will address the issue of the amount of training data for the SI model. Also, current ATRS system does not take advantage of various confidence scores available in leading recognition engines. We believe that using such confidence measures can improve the generation of semi-literal transcriptions considerably. We would also like to investigate the point at which the size of the various kinds of data used for adaptation stops making improvements in recognition accuracy.

## Acknowledgements

We would like to thank the anonymous reviewers of this paper for very helpful feedback. We thank Guergana Savova for excellent suggestions and enthusiastic support. We would also like to thank Jim Wu for valuable input.

## References

Digalakis, V and Neumyer, L. (1996). Speaker Adaptation Using Combined Transformation and Baysean Mehtods. IEEE Trans. Speech and Audio Processing.

Hazen, T and Glass, J (1997). A Comparison of Novel Techniques for Instantaneous Speaker Adaptation. In Proc. Eurospeech '97.

Heeman, P., Loken-Kim, K and Allen J. (1996). Combining the Detection and Correction of Speech Repairs. In Proc. ICSLP '96.

Huang, X. and Lee, K (1993). On Speaker – Independent, Speaker-Dependent, and Speaker-Adaptive Speech Recognition. In IEEE Transactions on Speech and Audio processing, Vol. 1, No. 2, pp. 150 – 157.

Legetter, C. and Woodland, P. (1996). Maximum Likelihood Linear Regression for Speaker Adaptation of Continuous Density HMM's. In Computer Speech and Language , 9, (171-186).

Pakhomov, S. (1999). Modeling Filled Pauses in Medical Transcriptions. In Student Section of Proc. ACL'99.

Pakhomov, S and Savova, G. (1999). Filled Pause Modeling in Quasi-Spontaneous Speech. In Proc. Disfluency in Spontaneous Speech Workshop at ICPHIS '99.

Savova, G, Schonwetter, M. and Pakhomov, S. (2000). Improving language model perplexity and recognition accuracy for medical dictations via within-domaininterpolation with literal and semi-literal corpora " In Proc. ICSLP '00.

Shriberg, E. 1994 Preliminaries to a Theory of Speech Disfluencies. Ph. D. thesis, University of California at Berkely.

Shriberg, E. and Stolcke, A. (1996). Word Predictability after Hesitations: A Corpus-based Study. In Proc. ICSLP '96.

Siu, M and Ostendorf, M. (1996). Modeling Disfluencies in Conversational Speech. In Proc. ICSLP '96.

Stolcke, A. and Shriberg, E. (1996). Statistical Language Modeling for Speech Disfluencies. In proc. ICASSP '96.

Stolcke A., Shriberg E., Bates R., Ostendorf M., Hakkani D., Plauche M., Tur G., and Lu Y. (1998). Automatic Detection of Sentence Boundaries and Disfluencies based on Recognized Words. Proc. Intl. Conf. on Spoken Language Processing.

Ström, N (1996): "Speaker Adaptation by Modeling the Speaker Variation in a Continuous Speech Recognition System," In Proc. ICSLP '96, Philadelphia, pp. 989-992.

Valtchev, V. Kershaw, D. and Odell, J. (1998). The Truetalk Transcriber Book. Entropic Cambridge Research Laboratory, Cambridge, England.

Wightman, C. W. and Harder T. A. (1999). Semi-Supervised Adaptation of Acoustic Models for Large-Volume Dictation" In Proc. Eurospeech '98. pp 1371-1374.

Weng, F., Stolcke, A., Sankar, A. (1997). Hub4 Language Modeling Using Domain Interpolation and Data Clustering. Proc. DARPA Speech Recognition Workshop, pp. 147-151, Chantilly, VA.