Flexible Transcription Alignment

Michael Finke and **Alex Waibel**
Interactive Systems Laboratories
Carnegie Mellon University
Pittsburgh, PA 15213 (USA)

Abstract - In this paper we present a set of techniques we employed in our Janus Recognition Toolkit (JRTk) Switchboard and CallHome recognizer in order to deal with imperfections in the transcriptions: inconsistent transcription of pronunciations and contractions as well as errors in utterance segmentations. These techniques consist of a dynamic, speaking mode dependent pronunciation model and a flexible utterance alignment procedure which is based on speaker adapted models (label boosting). The idea is (a) to automatically retranscribe the training corpus based on these models and procedures, (b) to train a recognizer based on these flexible transcription graphs and (c) to decode with a dynamic speaking mode dependent dictionary. The framework is successfully applied to increase the performance of our state-of-the-art JRTk Switchboard recognizer significantly.

1 Introduction

Recognition of conversational speech is one of the most challenging speech recognition tasks to-date. While recognition error rates of 10% or lower can now be reached on speech dictation tasks over vocabularies in excess of 60,000 words, recognition of conversational speech has persistently resisted most attempts at improvements by way of the proven techniques to-date. Difficulties arise from shorter words, telephone channel degradation, and highly disfluent and coarticulated speech.

We believe that the following inconsistencies in the transcriptions are a major source of problems when it comes to train and test on a large vocabulary conversational speech recognition corpus like Switchboard and CallHome:

- **Pronunciation Variations**: Spontaneous, conversational speech tends to be much more variable than the careful read speech that much of speech recognition work has focused on in the past. Pronunciation differences, in particular, represent one important source of variability that is not well accounted for by current recognition systems. For example, the word "BECAUSE" might be pronounced with a full or a reduced vowel in the initial syllable (IY vs. AX, respectively), or the whole initial syllable might be dropped (as in "CUZ"). These variations in pronunciation are not reflected in the word level transcriptions.
- **Crossword Pronunciaton Effects**: Contractions and reductions across word boundaries are especially hard to handle in state-of-the-art speech recognition engines. Words are typically the unit of training and recognition in speech recognizers. Even though allophonic modelling takes the neighbouring phones (also across word boundaries) into account, there are no means

so far that allow for reduction/rewriting of phones in a word depending on word context. Ignoring the word neighbours and still allowing for all sorts of phonetic reduction would result in a long list of confusion pairs of very frequent words. Consider of example word sequences like “KIND OF” and “SORT OF” which are often reduced to “KINDA” and “SORTA”. If, in order to capture this reduction of “OF”, we would introduce the pronunciation variant “OF(A)” transcribed with the unstressed vowel AX, the confusability in the dictionary would increase significantly.

- **Segmentation**: In Switchboard utterance boundaries are not well defined. It turned out that a lot of utterances were split incorrectly into utterances such that words at the beginning or end of an utterance were either only partially existent or not there at all.

The idea of this paper is to automatically retranscribe the training corpus using a speaking mode dependent pronunciation model which also takes crossword dependencies into account by merging words into multiword units. Training based on utterance transcription graphs which allow for a large number of alternative pronunciations and for a more flexible alignment and utterance segmentation yields a significant improvement in terms of word error rate of our Switchboard/CallHome recognizer.

2 Speaking Mode Dependent Pronunciation Modelling

In spontaneous conversational speech there is a large amount of variability due to accents, speaking styles and speaking rates (also known as the speaking mode) [4]. Because current recognition systems usually use only a relatively small number of pronunciation variants for the words in their dictionaries, the amount of variability that can be modelled is limited. Increasing the number of variants per dictionary entry is the obvious solution. Unfortunately, this also means increasing the confusability between the dictionary entries, and thus often leads to an actual performance decrease.

Similar to Tajchman *et al.* [5] we developed a probabilistic model based on context dependent phonetic rewrite rules (see Table 1) to come up with a list of possible pronunciations for all words or sequences of words [3]. In order to reduce the confusability of this expanded dictionary the idea is to annotate each variant of a word with an observation probability. To this aim we automatically retranscribe the corpus based on all the variants allowed. The alignments are then used to train a model of how likely which form of variation (i.e. rule) is and of the likelihood of a variant being observed in a certain context (acoustic, word, speaking mode or dialogue) is. For decoding, the probability of encountering pronunciation variants is then defined to be a function of the speaking style (phonetic context, linguistic context, speaking rate and durations). The probability function is learned through decision trees from rule based generated pronunciation variants as observed on the Switchboard corpus [3].

1	[AX IX] N → (E)N
2	[AX IX] M → (E)M
3	[AX IX] L → (E)L
4	[AX IX] R → AXR
5	[T D] → DX / [+VOWEL] __ [AX IX AXR]
6	[T D] R → DX
7	L → 0 / __ Y [AX IX AXR]
8	IY → Y / __ [AX IX AXR]
9	NG → N
10	HH → 0 / WB __
11	W → 0 / WB __
12	DH → 0 / WB __
13	[T D] → 0 / [+VOWEL] __ [TH DH]
14	[T D] → 0 / [+CONS +CONTINUANT] __ WB
15	R AX → ER / [-WB] __ [-WB]
16	T → 0 / [M N NG] __ [AX IX AXR]
17	BECAUSE → K [AH AO] Z
18	GOING TO → G AH N AX
19	WANT TO → W AH N AX
20	YOU KNOW → Y AX N OW
21	DO YOU → D Y UW

Table 1: **Pronunciation transformation rules** used in JRTk.

3 Multiwords and Multiword Clustering Algorithm

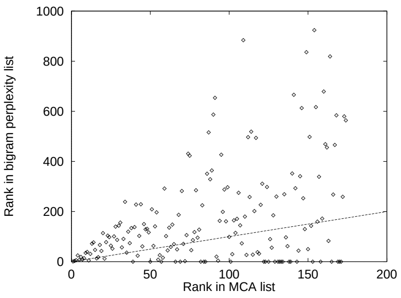
In order to model crossword pronunciation phenomena at least for very frequent sequences of words, we picked a list of about 200 so-called **multiwords** and added them to the dictionary. The criterion for combining words to multiwords was twofold: 1) mutual information between words, and 2) reduction in bigram perplexity (considering the multiword as a new language model token). It turns out that most of the multiwords consist of at least one of the short function words A, AND, AT, IT, OF or TO. The initial phonetic transcription of multiwords in the dictionary consisted of the concatenation of the transcriptions of the multiword's components.

Having multiwords in the dictionary, the question is how to treat these words in the decoding pass. We could either train our language model on a text file where sequences of words are replaced by multiwords or split multiwords when it comes to compute the LM probability for a given sequence of words. In [3] we presented evidence that on Switchboard and Callhome not modelling multiwords in the language model yields significantly better performance.

This raises the question on how we should pick the list of multiwords. It turns out that language model considerations (perplexity and/or mutual information) guided the search for multiwords even though we split multiword tokens with respect to the language model as described above. In the CLSP-WS97 at Johns Hopkins University we developed a new algorithm to find multiwords automatically (**Multiword Clustering Algorithm (MCA)**). This algorithm is no longer based on language modelling but on pronunciation modelling considerations instead. The goal is to iteratively find chunks of words which by merging them into a single token reduce the entropy of the pronunciation model most. The resulting list of multiwords differs significantly from the one found based on perplexity (see 1).

Multiword Clustering Algorithm (MCA)

1. Run acoustic alignment of the rule- or tree-based expanded dictionary to collect statistics on the distribution of variants per word.
2. Let $H(w) = -\sum_{v \in V(w)} p(v) \log p(v)$ be the entropy of the variants of word w.
3. For all word pairs compute the reduction in entropy by merging the pair into a multiword.
4. Pick the word pair which gives the best gain and replace all instances in the training data by the multiword.
5. Continue with 3.



MCA rank	Bigram rank	Word
1	1	YOU+KNOW
2	3	AND+UH
3	6	AND+I
4	25	GOING+TO
5	7	OF+THE
6	19	AND+THEN
7	9	KIND+OF
19	43	THE+UH
20	114	THAT+WE
31	156	AND+A
34	239	HAVE+BEEN
55	8	A+LOT
72	2	I+THINK
93	4	I+DON'T

Figure 1: **Rank Comparison** of the clustering algorithm derived multiwords and the rank based on a bigram perplexity reduction criterion. It turns out that for some of the multiwords the two rankings are quite different: On the one hand there is a list of MCA based multiwords that have a particularly high ranking compared to the bigram based list. This is because there is a significant word pair dependent pronunciation effect. On the other hand there are word pairs like "A LOT" that score pretty good in terms of the bigram criterion but since there is no significant context dependent pronunciation variation involved, MCA ranks them lower.

4 Flexible Alignment of Transcription Graphs

4.1 Utterance Transcription Graphs

In order to train our speech recognizer based on unreliable transcriptions we implemented a **Flexible Transcription Alignment (FTA)** procedure in JRTk [1, 3]. Instead of aligning the plain transcription of an utterance we generate a hidden markov model for each utterance that allows for

1. all alternative pronunciations in the dictionary for each word,

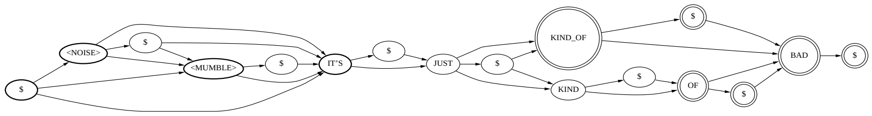


Figure 2: **Utterance Transcription Graph** of a Switchboard utterance ``IT-IT'S JUST KIND OF BAD''. Bold circles are potential utterance initial states and double circles mark final states. $ represents the silence model.

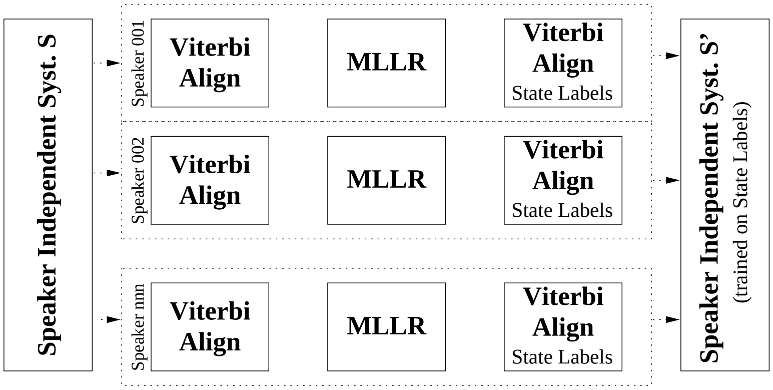

Speaker Independent Syst. S
Speaker 001
Viterbi Align
MLLR
Viterbi Align State Labels
Speaker 002
Viterbi Align
MLLR
Viterbi Align State Labels
Speaker nnn
Viterbi Align
MLLR
Viterbi Align State Labels
Speaker Independent Syst. S'
(trained on State Labels)

Figure 3: **Label Boosting**: Speaker adapted models are used in order to run the forced alignment of the utterance transcription graphs and thus find the most likely transcription (state labels).

2. multiwords as alternative word to the sequence of words they consist of,
3. beginning and ending words of an utterance being optional,
4. optional silence or breathing models between words,
5. optional noise words to start or end an utterance.

See Figure 2 for an example of such an utterance transcription graph.

4.2 Label Boosting

Instead of relying on a speaker independent acoustic model to align the flexible utterance HMMs, we adapt the recognizer using maximum likelihood linear regression (MLLR) to derive a speaker dependent recognizer for each speaker. The speaker dependent forced alignment is then used to determine a new transcription of the training corpus in terms of pronunciation variation and utterance segmentation (see Figure 3) [6, 1].

Table 2 shows the resulting alignment for a Switchboard utterance. The underlined words were part of the original Switchboard transcription. Parentheses mark pronunciation variants with the rule numbers that they were derived from attached. In this sample utterance we observe among other things, that the GOING TO goes to GONNA rule was applied, that the ending NG in the word ASKING is reduced to N (rule 9) and that KIND OF surfaced as KINDA.

$(<BREATH>) <NOISE>(BREATH) $ AND $(<SBREATH>) I $ YOU_KNOW $ IT'S $ I GUESS IT'S SO NORMAL TO(2) $(<BREATH>) START TO WONDER $ ABOUT THAT EVEN IF(2) SHE DOESN'T(2) NEED THAT BUT $(<SBREATH>) YOU_KNOW SHE'S KIND_OF(KINDA/1) ASKING(1/9/9) QUESTIONS ABOUT WHAT(2) $(<BREATH>) WELL WHAT'S GOING_TO(GONNA/1) HAPPEN THIS CAN'T LAST FOREVER(1/4,18,20/18,20) AND $(<SBREATH>) <NOISE>(THROAT)

Table 2: **FTA transcript** of a Switchboard utterance; parentheses mark pronunciation variants and $ is the silence word.

5 Results

The test set to evaluate the use of the flexible transcription alignment approach consists of the Switchboard and CallHome partitions of the 1996 NIST Hub-5e evaluation set. All test runs used the JRTk Switchboard recognizer[2].

The preprocessing of the system consists of extracting an MFCC based feature vector every 10 ms. The final feature vector is computed by a truncated LDA transformation of a concatenation of MFCCs and their first and second order derivatives. Vocal tract length normalization and cepstral mean subtraction are used to extenuate speaker and channel differences.

The rule-based expanded dictionary that was used in these tests included 1.78 pronunciations variants/word, compared to 1.13 for the baseform dictionary (PronLex). The first list of results in Table 3 is based on a recognizer whose polyphonic decision trees were still trained on viterbi alignments based on the unexpanded dictionary. We compare a baseline system trained on the base dictionary with an expanded dictionary FTA trained system tested in two different ways: with the base dictionary and with the expanded one. It turns out, that FTA training reduces the word error rate significantly, which means, that we improved the quality of the transcriptions through FTA and pronunciation modelling. Due to the added confusability of the expanded dictionary the test with the large dictionary without any weighting of the variants yields slightly worse results than testing with the baseline dictionary.

Condition	**SWB WER**	**CH WER**
Baseline	32.2%	43.7%
FTA training/test w.basedict	30.7%	41.9%
FTA training/test w.exp.dict	31.1%	42.5%

Table 3: Recognition results using flexible transcription alignment training and label boosting. The test using the expanded dictionary was done without weighting the variants.

Adding vowel stress related questions to the phonetic clustering procedure and regrowing the polyphonic decision tree based on FTA labels improved the performance by 2.6% absolute on SWB and 2.2% absolute on CallHome. Table 4 shows results for mode dependent pronunciation weighting. We gain about an additional 2% absolute by weighting the pronunciation based on mode related features.

Condition	**SWB WER**	**CH WER**
unweighted	28.7%	38.6%
weighted $p(r\|w)$	27.1%	36.7%
weighted $p(r\|w, m)$	26.7%	36.1%

Table 4: Results using different pronunciation variant weighting schemes.

6 Conclusion

We presented an approach to deal with imperfect word level transcriptions when it comes to training a speech recognition system. A pronunciation model was defined to incorporate speaking style related information into the probability estimates for different pronunciation variants. Preliminary results show a significant increase in word accuracy through flexible transcription alignment, label boosting and using a probability weighted pronunciation dictionary within the JRTk Switchboard recognizer. The JRTk recognizer based on speaking mode dependent pronunciation modelling as presented here was one of the two winning systems of the 1997 NIST Hub5-e evaluation and thus proved to be state-of-the-art.

References

[1] M. Finke, J. Fritsch, P. Geutner, K. Ries, T. Zeppenfeld, and A. Waibel. The JanusRTk Switchboard/Callhome 1997 Evaluation System. In *Proceedings of LVCSR Hub 5-e Workshop*, May 1997.

[2] Michael Finke. The JanusRTk Switchboard/Callhome 1997 Evaluation System: Pronunciation Modeling. In *Proceedings of LVCSR Hub 5-e Workshop*, May 1997.

[3] Michael Finke and Alex Waibel. Speaking Mode Dependent Pronunciation Modeling in Large Vocabulary Conversational Speech Recognition. In *Proceedings of Eurospeech-97*, September 1997.

[4] M. Ostendorf, B. Byrne, M. Bacchiani, M. Finke, A. Gunawardana, K. Ross, S. Roweis, E. Shriberg, D. Talkin, A.Waibel, B. Wheatley, and T. Zeppenfeld. Systematic Variations in Pronunciation via a Language-Dependent Hidden Speaking Mode. In *International Conference on Spoken Language Processing*, Philadelphia, USA, 1996.

[5] G. Tajchman, E. Fossler, and D. Jurafsky. Building Multiple Pronunciation Models for Novel Words using Exploratory Computational Phonology. In *Proceedings Eurospeech*, pages 2247–2250, 1995.

[6] Torsten Zeppenfeld, Michael Finke, Klaus Ries, Martin Westphal, and Alex Waibel. Recognition of Conversational Telephone Speech using the JANUS Speech Engine. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, Munich, Germany, 1997. IEEE.