Error Signatures to identify Errors in ASR in an unsupervised fashion

Dominic Telaar, Jochen Weiner, Tanja Schultz

Cognitive Systems Lab
Karlsruhe Institute of Technology (KIT), Karlsruhe, Germany
University of Bremen, Bremen, Germany
dominic.telaar@kit.edu

Abstract

Large scale ASR systems are trained on thousands of hours of speech. Usually, many of these training data were automatically transcribed by another ASR system due to a lack of manual transcriptions and a lack of resources to transcribe them. Systems trained in such a fashion are biased towards the transcription system. In the past, confidence models have been investigated to exclude data from training. We propose to investigate areas of low confidence by extending our previous work. For this purpose we aggregate potential errors of ASR systems by ascribing a list of attributes to each potential error and find a set of attributes which best describe the errors encountered on an automatically transcribed set. We call these characteristic sets of attributes *Error Signatures*. Examples of attributes are word identity, phonemes, acoustic models, word context, speaker id, and language id. For each *Error Signatures*, an error ratio is computed, giving the probability that the signature properly describes the error. Error ratios and occurrence frequencies are used to sort the signatures and present them to an expert to fix the *Error Signatures* underlying shortcomings of the ASR system.

1. Introduction

Much of the previous work on the topic of system analysis and error assignment in automatic speech recognition needs supervision in form of reference transcriptions to conduct their investigation. Earlier work by Chase [1] and Nanjo et al. [2] focused on sentence by sentence error assignment given an ASR hypothesis and reference. We investigate an approach to error analysis which does not require reference transcriptions. Most large scale ASR systems trained today are based on thousands of hours of speech without reference transcriptions, employing confidence models to exclude potentially erroneous segments from the training data, keeping the systems from deteriorating. In this work we want to analyze regions of low confidence, group them together and report on commonalities, without relying on reference transcriptions. The approach can be integrated into an iterative training scheme: after finishing a training iteration the untranscribed training data is decoded with the latest ASR system and an error profile with *error signatures* is generated. The error profile can then be consulted by an expert to pinpoint shortcomings and improve the ASR system.

2. Error signatures

This section introduces the notion of *error signatures*, how to obtain them, and how to use them to identify causes of errors in an ASR system. Each signature consists of a set of attribute-value pairs to identify potentially problematic words, examples of attributes are:

- word identity
- phonemes
- acoustic models
- word context
- word confidence score
- language id/language id context

Additional attributes, such as signal-to-noise ratio or part-of-speech tags, can be added on the fly to our framework, by either extending the dictionary and adding additional word-level attributes or supplying sentence-level attributes during the collection of statistics on the different attributes, as described in Section 2.1.

The process of finding *error signatures* is similar to concept discovery in data mining [3]. In data mining concepts are discovered by finding patterns of jointly occurring attributes. Transferring this concept to an ASR hypothesis, we want to find a set of attribute-value pairs which jointly appears in low confidence words and is only rarely found in words with high confidence values. We refer to such a set of attribute-value pairs as an *error signature*.

We call the procedure of finding *error signatures* on a data set *error profiling*. The process of obtaining the *error signatures* is carried out in two steps. In the first step, as described by Evermann et al. [4], and attribute-value pairs are confidence scores are computed for each word in the 1-best hypothesis of each utterance. We assume that low confidence scores are indicative of an error. In a subsequent step, attribute-value pairs are clustered to *error signatures*.

2.1. Collecting statistics of attributes

The input to our framework is the 1-best hypothesis of all sentences in the training or development set. Since no reference transcription is available as in the work by Chase [1], we cannot derive error regions from the hypothesis and have to rely on word confidence scores extracted from the decoding lattice. After extracting word confidence scores for each word in a hypothesis, additional attribute-value pairs are assigned to it.

Table 1 gives an excerpt of the attribute-value pairs assigned to the hypothesis segment ”都是” from the segment ”都 都是 SIL” in the SEAME corpus [5]. The five phoneme attributes */d/ /o/ /w/ /S/ /i/* were added, corresponding to the segments pronunciation in the dictionary. The segment is preceded by a Mandarin word (-1=MAN) and specifically preceded by the segment ”都” (-1=都). Furthermore, the word is assigned the language id ”MAN”..

Table 1: *Attributes for "都是" from a segment in the SEAME corpus ("都 都是 SIL").*.

Category	Attributes
Confidence	0.4
Token	都是
Phonemes	/d/ /o/ /w/ /S/ /i/
Context	-1=MAN -1=都 +1=SIL
Language Id	MAN

2.2. Computing error signatures

After all utterances are processed and attributes have been assigned to each word in the hypothesis, we employ a bottom-up greedy clustering to find the *error signatures* which best explain the errors we encounter in our data set. A prerequisite for the clustering process is the ability to compare two *error signatures* and decide if one signature is a better fit for the potential errors than the other.

In data mining two measures are used for this purpose: a concept's support and confidence. Support is the number of hypothesis words matching the *error signature*. We derive the confidence in an *error signature* by computing the average error probability of all matching hypothesis words, where the error probability for any given word is derived from its confidence score. We refer to this measure as the *error ratio*. It can be interpreted as the probability that a hypothesis word with that signature was misrecognized.

The clustering algorithm we employ to find the *error signatures* uses only discrete attributes, testing for the presence or absence of an attribute-value pair. In order to make use of continuous-valued attributes, such as word confidence score, the value range of these attributes has to be discretized.

We chose to use the minimum entropy partitioning algorithm introduced by Fayyad et al. [6]. This greedy algorithm recursively chooses the partition of the value range with the biggest decrease in entropy. The algorithm starts with one partition containing the complete value range and proceeds partitioning until either the information gain of an additional partition or the occupancy count of the resulting partitions fall below a threshold.

Initially we start the clustering process with one *error signature* for every attribute-value pair whose support exceeds a threshold. New *error signatures* are created by recursively expanding the list of attribute-values of previous *error signatures* with an additional attribute-value pair. Criteria for these newly created *error signatures* are:

- an *error signature* must cover a minimum number of error words (min. support)
- probability of *error signatures* indicating an error has to exceed a certain threshold (min. *error ratio*)
- there is no *error signature* which consists of a subset of attribute-values to the *error signature* in question and has a higher *error ratio*
- there is no *error signature* which represents the same group of words and has a higher *error ratio* (removal of dominated *error signatures*)

The clustering process stops if no *error signature* can be expanded by an additional attribute-value pair without violating one of these criteria. To speed up the clustering process the attribute-value pairs which are considered next as a potential addition to an existing signature are sorted in ascending order of frequency. Processed in such a way attribute-value pairs with low frequency count can be processed quickly since their *error signatures* quickly fall below the minimum support requirement and can be disregarded from other *error signatures* since any *error signature* not violating any of the conditions should have been found while processing the attribute-value pair in the first place.

The algorithm to obtain the *error signatures* is implemented in our in-house toolkit BioKIT [7] which we also used to obtain the decoding results. Examples of *error signatures* are presented in our experiments on the SEAME corpus and ILSE corpus in Section 5.

3. Error Correction

After *error signatures* have been computed it is up to the expert to scan the list of signatures, and deduce which component of the system could be responsible for producing the observed hypotheses. Furthermore, the expert has to conclude if the problem can be fixed my modifying a component of the ASR system and if that is the case which changes are most likely to fix the encountered problem.

To keep the required amount of manual work to a minimum, the *error signatures* presented to the expert have to be limited to a select few meaningful signatures. Therefore, we employed heuristics to filter the computed *error signatures*. The data for the figures in this section are derived from the 19 hour baseline system we trained on the SEAME corpus, which is presented in Section 5.1. *Error signatures* were computed on the remaining 39 hours of speech from the original SEAME corpus training set, as described in Section 4. The total number of *error signatures* without filtering, with the exception of a minimum number of samples per word as defined in the *error signature* algorithm in Section 2, is $31,332$.

3.1. Error Ratio filter

Figure 1 depicts all *error signatures* and their corresponding *error ratio*, as they were found on the 39 hours of the SEAME corpus training set, set aside for computing *error signatures*. It shows that there are only a few *error signatures* who are worth looking into. Any *error signature* which has an *error ratio* below or equal to 0.5 will potentially introduce more errors than fix existing ones by trying to remove its underlying cause of error. By applying a threshold of 0.5 to the error ratio of each error signature, the number of error signatures which are of interest to an expert for error correction are reduced from $31,223$ to 521.

3.2. Frequency filter

Since the number of error signatures that can be viewed by an expert should be limited to a select meaningful few, the potential impact of each error signature on the error rate of the system should be as big as possible. Thus, any error signature selected for further investigation has to represent a minimum number of matching words of the data set being analyzed. The total number of error words in the 39 hour data set amounts to $500,051$ distributed over $30,294$ utterances. Limiting the minimum number of affected utterances to at least 0.3% of the 39 hours would require a minimum number of matching words for each *error signature* to at least 91, reducing the $31,223$ *error signatures* to $22,927$, see Figure 2. In combination with the

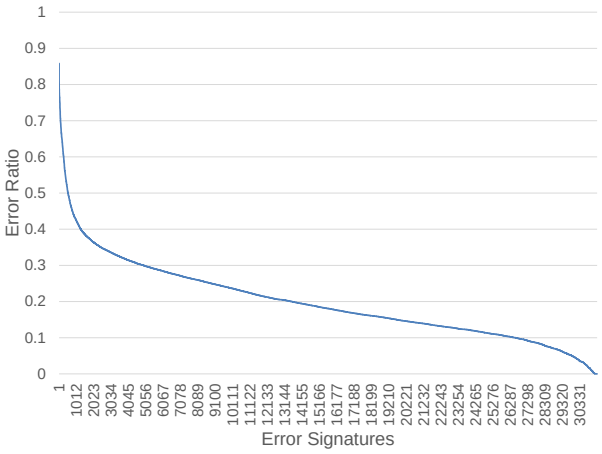

Figure 1: Error signatures on 39 hours portion of SEAME training corpus sorted by their error ratio.

error ratio filter from Section 3.1, the number of viable *error signatures* is reduced to 150 or 0.5% of the original *error signatures*.

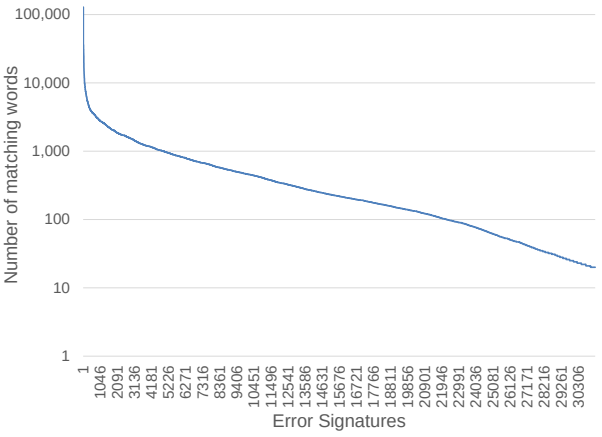

Figure 2: Error signatures on 39 hours portion of unsupervised SEAME training corpus sorted by the number of matching words. Y-axis is in logarithmic scale.

4. Database

We applied the proposed error signature algorithm to the two corpora described in this section.

4.1. SEAME corpus

The code-switching speech corpus we used is called SEAME (South East Asia Mandarin-English). It has been recorded in Singapore and Malaysia by Lyu et al. [5] and contains spontaneously spoken interviews and conversations. Originally, it was used for the research project "Code-Switch"which was jointly performed by Nanyang Technological University (NTU) and Karlsruhe Institute of Technology (KIT). The corpus contains about 62 hours of audio data and manual transcriptions which are separated into training, development and evaluation set (refer to Table 2).

The words can be categorized into four language classes: Mandarin (58.6% of all tokens), English (34.4% of all tokens), particles (Singaporean and Malayan discourse particles, 6.8% of all tokens) and other languages (0.4% of all tokens). The average number of code-switching events between Mandarin and English is 2.6 per utterance. The average duration of monolingual English and Mandarin segments is only 0.67 seconds and 0.81 seconds, respectively. In total, the corpus includes 9,210 unique English and 7,471 unique Mandarin words.

Table 2: *The SEAME database.*

	Train set	Dev set	Eval set
#Speakers	157	8	8
#Tokens	681k	28k	14k
Duration (hrs)	58.4	2.5	1.1

To extract *error signatures* in an unsupervised fashion we further split the SEAME training set into two parts. The set was split such that the ratio of female to male, Malayan to Singaporean speakers, and conversational style to interview style recordings are kept roughly the same. The first portion of 19 hours is used to train the baseline acoustic and language model the second, larger portion is used to extract the *error signatures* and perform unsupervised acoustic and language model training. Details of the resulting two parts are given in Table 3.

Table 3: *Separation of SEAME training data into portions for supervised and unsupervised training.*

Set	#Speakers	#Tokens	Duration (hrs)
Supervised	58	235k	19.1
Unsupervised	99	446k	39.3
Total	157	681k	58.4

4.2. ILSE corpus

For the ILSE study [8, 9], an interdisciplinary team at Heidelberg University was given the task of collecting a large-scale investigation into the impact of aging into several aspects of daily life. During the course of 20 years they conducted interviews and medical examinations on about 1,000 participants. The collected German speech corpus was partly digitized and transcribed. Details of the corpus and the sets we divided the corpus into is given in Table 4.

Table 4: *The ILSE corpus.*

	Training	Development	Evaluation
#speakers	68	10	14
#Tokens	2,000k	270k	500k
Duration (hrs)	265.4	35.9	62.0

5. Experimental setup and results

In our experiments we focused on the above two corpora to show our approach of error profiling. To demonstrate the usefulness of our approach we compare the error rates using unsupervised acoustic model training and language model adaptation based on confidences against unsupervised training with implementation of error fixes derived from the found error signatures.

5.1. SEAME corpus

5.1.1. System setup

Recognition results were obtained in a multi-pass system. The last pass is a deep neural network (DNN) acoustic model trained with the Kaldi toolkit [10] using an fMLLR adaptation from the previous pass. The fMLLR transforms were created in a first pass by using Kaldi trained Gaussian Mixture Models. The input for the DNN is a 440-dimensional feature vector consisting of 11 stacked 40 dimensional LDA transformed stacked MFCC feature vectors.

The DNN is trained by first performing pre-training, followed by 13 iterations optimizing cross-entropy and finally four iterations of state-level minimum Bayes risk (sMBR) sequence training [11]. The neural network consists of 6 hidden layers with 2,048 nodes each and an output layer with 3,194 nodes.

The system is trained only on the 19 hours of the supervised training set. The baseline language model is a Kneser-Ney trigram estimated on the supervised training set's text and monolingual English and Mandarin texts taken from NIST using the SRILM toolkit [12]. The perplexity on the SEAME development set is 349.6. The perplexity on the unsupervised training set is 413.9. Decoding results were obtained using BioKIT [7].

5.1.2. Baseline

In a first step, the initial system trained on 19 hours of speech is used to decode the unsupervised training set of 39 hours. The mixed error rate on that unsupervised training set is 37.85%. Figure 3 shows the confidence distribution in deciles over the unsupervised training set.

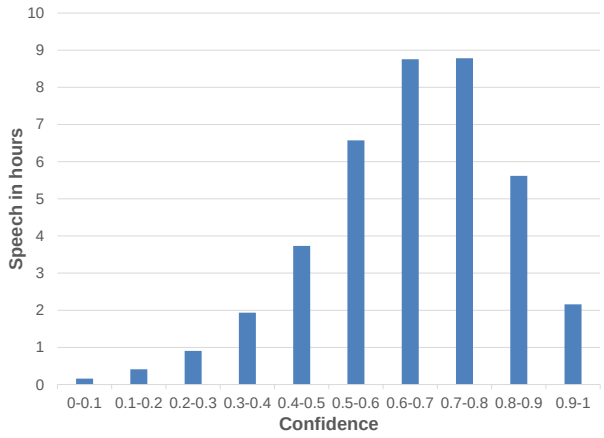

Figure 3: Unsupervised training data split into decile confidence bins after decoding and confidence estimation.

5.1.3. Finding error signatures

We apply the error signature algorithm to the unsupervised training set and filter the resulting error signatures according to the steps described in Section 3. From the resulting 150 error signatures we pick the error signatures with the highest *error ratios* for further analysis. Chosen *error signatures* are shown in Table 5 and the following list describes the steps we took to improve the model, based on these *error signatures*:

- Signature #1: Investigating the dictionary it was striking that pronunciations for "like" and "kite" where "/l/ /ay/" and "/k/ /ay/" respectively. The cause for these faulty pronunciations was due to an error in the automatic generation of pronunciations which also affected other words in the vocabulary, which ended in plosives such as /t/, /k/, and /p/. Two clean the dictionary for decoding in a first step only pronunciations of English words were kept which are already present in the training dictionary. In a second step, frequencies of all words that appeared during decoding on the 39 hour unsupervised training set were counted. Afterwards, we counted the total number of appearances for each word and, after scanning the list, set a threshold of 15% of how frequently each pronunciation had to appear to remain in the final decoding dictionary. The number of entries in the dictionary was reduced from 52,225 to 35,645 while keeping the vocabulary size constant.
- Signature #2-#3: No specific cause could be identified for signatures #2 and #3. Sentences matching #3 were excluded from unsupervised acoustic model training. Sentences matching signature #2 were shortened to exclude the last two words.

To estimate the impact of our fixes we compare two systems, one trained with utterances from the unsupervised training set whose confidence score exceeds 0.8 [13], the other system additionally using the fixes derived from the *error signatures*. Mixed error rates of the adapted systems and baseline are presented in Table 6. The impact of selecting different amounts of data for unsupervised acoustic modeling and language modeling adaptation based on signatures #2-#3 led to differences in performance on the development set. These differences were not significant on the development set at a significance level of 0.05. Only changing the decoding dictionary had a noticeable impact on system performance. By applying the fixes to the dictionary we achieved a significant reduction in MER ($p = 0.007$) from 29.1% on the adapted baseline system to 28.3% when using error signatures.

5.2. ILSE corpus

In this section we present the experiments conducted on the ILSE corpus and the improvements we achieved by applying fixes to errors found with the *error signatures*. In the previous experiment on SEAME the computation of the *error ratio* was directly derived from the words' confidence scores matching the signature in question. For the ILSE corpus we adapted the *error ratio* computation by computing the minimum confidence score of each attribute-value pair for each word. The reason is that in this way even if the confidence score of a particular word is high, low confidence phonemes or acoustic models within that word can be identified.

5.2.1. System setup

The acoustic model is a deep neural network with the same preprocessing as described in Section 5.1.1. Since no speaker information is available in the interview files, the fMLLR transform is omitted, resulting in a one pass system. The dictionary is build based on the transcriptions in the training data and contains 74,423 entries, the size of the vocabulary is 71,889.

5.2.2. Baseline

Due to the long audio files of up to 45 minutes and their accompanying transcriptions being erroneous and not verbatim with

Table 5: *Examples for* error signatures *found on the unsupervised training set of the SEAME corpus. Word on the right hand side of word confusion's column is the word in the 1-best hypothesis.*

Sig.#	Error-Ratio	#Occ.	Error-Signature	Word confusions
1	0.79	124	+1=</s> AM no. 274	的 → kite direct → like
2	0.76	147	+2=</s> 0="而已"	yen → "而已" "来了" → "而已"
3	0.75	382	min. duration [noise]	"现在" → [noise] center → [noise]

Table 6: *Mixed error rates on evaluation (development) set, comparing baseline system with system using error signatures on the SEAME corpus.*

System	Baseline	Confidence ≥ 0.8	Confidence ≥ 0.8 and error signatures
no adapt	30.36% (36.22%)	-	-
AM adapt	-	29.25% (34.88%)	29.10% (35.28%)
AM & LM adapt	-	29.09% (34.53%)	28.86% (34.85%)
AM & LM & dict adapt	-	-	**28.29% (34.34%)**

no timestamps, forced alignment cannot be applied successfully to train an initial system. Therefore, the initial system we perform our error signature algorithm on was trained on a subset of the 265 hours of training data. Using a long audio alignment procedure we extracted segments from the audio whose transcriptions are correct with a high confidence. Ultimately, the system we employ is trained on 44 hours of training data. The language model used for decoding is build on the reference transcriptions of the complete training set.

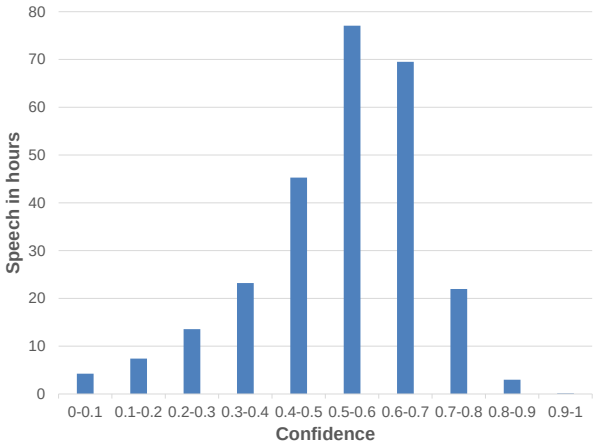

Figure 4: Training data split into decile confidence bins after decoding and confidence estimation.

5.2.3. *Finding error signatures*

Since the condition of the manual transcriptions of the training set is such that they only loosely match the actual content, we chose to conduct our experiments on the complete 265 hour training set. Due to the extensive size of both development and test set we only report word error rates on a subset of both development and evaluation set. From the original development and evaluation sets, we selected the first 45 minutes of each interview as development and evaluation set.

Table 7 shows the error signatures we found and implemented fixes for. The fixed *error signatures* were among the 70 most promising signatures, 25 of which where concerned with the noise acoustic model. The following list shows the fixes we implemented to remedy the problems identified by the *error signatures*:

- Signature #1: The acoustic model with id 140 models the phoneme /n/ and frequent potential confusions of signature #1 are with the word "und" which should only rarely appear at sentence ends. Investigating the language model probabilities for the bigram "und </s>", turned out that this bigram is far more likely than most other bigrams with the word "und". Further looking at the training text used for estimating the language model probabilities, we found that the text contains a sizeable number of partial sentences. To remedy this situation we removed sentence end tags from all sentences ending in the word "und". The language model was then re-estimated.
- Signature #2-#5: The noise token frequently appears in areas of low confidence. We noticed that signatures regarding the noise token contained more than the expected five acoustic models for the five state noise HMM. Further investigation of the acoustic model turned out that the noise model had erroneously been trained as a context-dependent model. In the new model we retrained noise as a context-independent model.
- Signature #6: This signature mostly concerns the word "haben" with only one pronunciation in the dictionary: "/h/ /a/ /b/ /etu/ /n/". The last two phonemes are minimum duration only (duration of three frames), which indicates that the actual pronunciation contains less phonemes. As a solution we added another colloquial pronunciation of "haben" to the dictionary.
- Signature #7: The acoustic model with id 2805 in the *error signature* models the phoneme /r/. The signature shows that the word "oder", which is quite frequent in the German language, is often confused with other words. The dictionary contained two pronunciations for this word, the first one being "/ol/ /d/ /atu/" and the second one "/ol/ /d/ /etu/ /r/". Since the second pronunciation is rather unlikely of being correct it was removed from the dictionary.
- Signature #8-#9: The last two signatures are words whose pronunciations either end in "/r/ /etu/ /n/" or "/d/ /etu/ /n/". As a fix to this problem we added additional pronunciations to the dictionary which more closely match the actual pronunciation by omitting the phoneme /etu/.

Table 7: *Examples for* error signatures *found on the training set of the ILSE corpus. Word on the right hand side of word confusion's column is the word in the 1-best hypothesis.*

Sig.#	Error-Ratio	#Occ.	Error signature		Word confusions	
1	0.99	1,251	+1=</s>	AM no. 140	die → und	möbeln → und
2	0.98	1,149		AM no. 2542		SIL → [noise]
3	0.98	773	AM no. 2673	AM no. 850		[redacted] → [noise]
4	0.98	10,038	AM no. 349	AM no. 4531	beste → [noise]	bleibe → [noise]
5	0.97	2,381		AM no. 4458		[redacted] → [noise]
6	0.96	1,240	0=/h/	AM no. 277	[unk] → haben	daheim → haben
			min. duration /etu/	min. duration /n/	oben → haben	feilen → haben
7	0.94	634	0=oder	AM no. 2805	[redacted] → oder	tätigkeit → oder
8	0.94	763	AM no. 1244	AM no. 3098	drin → besonderen	marschiert → marschieren
			AM no. 4360	min. duration /r/	überall → waren	das → verfügbaren
9	0.94	1,902	AM no. 1334	AM no. 3098	hatten → werden	freunden → geworden
			AM no. 697	min. duration /etu/	gut → wurden	zum → worden

Table 8: *Mixed error rates on evaluation (development) set, comparing baseline system with system using error signatures on the ILSE corpus.*

System	Baseline	Confidence ≥ 0.8	Confidence ≥ 0.8 and error signatures
no adapt	64.82% (67.11%)	-	-
AM adapt	-	64.02% (66.54%)	63.98% (66.05%)
AM & LM fix	-	-	63.98% (65.77%)
AM & LM & dict fixes	-	-	**63.68% (65.50%)**

Similar to our experiments on the SEAME corpus we trained two new systems based on the utterance-level confidence scores depicted in Figure 4. The total amount of training data for the new models is 47 hours. Word error rates for the baseline system and our improved systems are shown in Table 8. Since the additional amount of selected training data was so small, the language model was not adapted. Even though the acoustic model and language model fixes had an impact on the development set, the evaluation set performance increase was marginal. Only the changes to the dictionary consistently improved the system performance. In conclusion we achieved a significant ($p = 0.012$) word error rate reduction of 0.53% relative.

6. Conclusion

We introduced an algorithm for error analysis, which based on the lattices of an ASR system, groups low confidence words together using various attributes. We showed that the resulting *error signatures* were useful to improve ASR systems build on the SEAME code-switching corpus and the ILSE corpus. We were able to significantly improve the error rate by 2.75% and 0.53% respectively relative compared to models trained with unsupervised acoustic model training only. As future work we plan to apply the algorithm iteratively to see if additional errors can be identified and how many iterations are needed until no more errors can be found.

7. References

[1] L. L. Chase, "Error-Responsive Feedback Mechanisms for Speech Recognizers," Ph.D. dissertation, Pittsburgh, PA: Carnegie Mellon University, 1997.

[2] H. Nanjo, A. Ri, and T. Kawahara, "Automatic Diagnosis of Recognition Errors in Large Vocabulary Continuous Speech Recognition System," *Joho Shori Gakkai Kenkyu Hokoku*, vol. 99, no. 64, pp. 41–48, 1999.

[3] P. Valtchev, R. Missaoui, and R. Godin, "Formal concept analysis for knowledge discovery and data mining: The new challenges," in *Concept lattices*. Springer, 2004, pp. 352–371.

[4] G. Evermann and P. C. Woodland, "Large vocabulary decoding and confidence estimation using word posterior probabilities," in *ICASSP*, vol. 3, 2000, pp. 1655–1658.

[5] D.-C. Lyu, T.-P. Tan, E.-S. Chng, and H. Li, "An analysis of a Mandarin-English code-switching speech corpus: SEAME," *Age*, vol. 21, pp. 25–33, 2010.

[6] U. Fayyad and K. Irani, "Multi-interval discretization of continuous-valued attributes for classification learning," *Proc. IJCAI93, Chambery, France, Aug.*, pp. 1022–1027, 1993.

[7] D. Telaar et al., "Biokit - real-time decoder for biosignal processing," in *Interspeech*, 2014.

[8] P. Martin and M. Martin, "Design und Methodik der Interdisziplinären Längsschnittstudie des Erwachsenenalters," in *Aspekte der Entwicklung im mittleren und höheren Lebensalter: Ergebnisse der Interdisziplinären Längsschnittstudie des Erwachsenenalters (ILSE)*, P. Martin, K. U. Ettrich, U. Lehr, D. Roether, M. Martin, and A. Fischer-Cyrulies, Eds. Steinkopff, 2000, pp. 17–27.

[9] U. Lehr, H. Thomae, M. Schmitt, and E. Minnemann, "Interdisziplinäre Längsschnittstudie des Erwachsenenalters: Geschichte, theoretische Begründung und ausgewählte Ergebnisse des 1. Messzeitpunktes," in *Aspekte der Entwicklung im mittleren und höheren Lebensalter: Ergebnisse der Interdisziplinären Längsschnittstudie des Erwachsenenalters (ILSE)*, P. Martin, K. U. Ettrich, U. Lehr, D. Roether, M. Martin, and A. Fischer-Cyrulies, Eds. Steinkopff, 2000, pp. 1–16.

[10] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, and P. Schwarz, "The Kaldi Speech Recognition Toolkit," in *IEEE Automatic Speech Recognition and Understanding (ASRU) Workshop*, 2011, pp. 1–4.

[11] K. Veselỳ, A. Ghoshal, L. Burget, and D. Povey, "Sequence-discriminative training of deep neural networks." in *INTERSPEECH*, 2013, pp. 2345–2349.

[12] A. Stolcke, "SRILM —An Extensible Language Modeling Toolkit," in *Interspeech*, 2002, pp. 901–904.

[13] A. Laurent, W. Hartmann, and L. Lamel, "Unsupervised acoustic model training for the korean language," in *Chinese Spoken Language Processing (ISCSLP), 2014 9th International Symposium on*. IEEE, 2014, pp. 469–473.