Using Adaptation to Improve Speech Transcription Alignment in Noisy and Reverberant Environments

Y. Mamiya[1], *A. Stan*[2], *J. Yamagishi*[1,3], *P. Bell*[1], *O. Watts*[1], *R.A.J. Clark*[1], *S. King*[1]

[1]Centre for Speech Technology Research, University of Edinburgh, United Kingdom
[2]Department of Communications, Technical University of Cluj-Napoca, Romania
[3]National Institute of Informatics, Tokyo, Japan

{yoshitaka.mamiya,jyamagis,owatts}@inf.ed.ac.uk, robert@cstr.ed.ac.uk
adriana.stan@com.utcluj.ro, {Peter.Bell,Simon.King}@ed.ac.uk

Abstract

When using data retrieved from the internet to create new speech databases, the recording conditions can often be highly variable within and between sessions. This variance influences the overall performance of any automatic speech and text alignment techniques used to process this data. In this paper we discuss the use of speaker adaptation methods to address this issue. Starting from a baseline system for automatic sentence-level segmentation and speech and text alignment based on GMMs and grapheme HMMs, respectively, we employ Maximum A Posteriori (MAP) and Constrained Maximum Likelihood Linear Regression (CMLLR) techniques to model the variation in the data in order to increase the amount of confidently aligned speech. We tested 29 different scenarios, which include reverberation, 8 talker babble noise and white noise, each in various combinations and SNRs. Results show that the MAP-based segmentation's performance is very much influenced by the noise type, as well as the presence or absence of reverberation. On the other hand, the CMLLR adaptation of the acoustic models gives an average 20% increase in the aligned data percentage for the majority of the studied scenarios.

Index Terms: speech alignment, speech segmentation, adaptive training, CMLLR, MAP, VAD

1. Introduction

For any corpus-based speech synthesis system or automatic speech recognition system, one of the most important considerations is the selection of high quality speech data for training purposes. For a limited number of languages, such as English, Spanish, French, and German, developers can chose from many widely available specifically-prepared resources. However, for most of the world's languages such speech databases are not readily available. Even for apparently well-resourced languages, the specific content of available data might not be suitable for a particular need – for example data in a sports news speaking style would probably not be as readily available as broadcast news data. In such situations, new resources either need to be recorded from scratch, or created from existing sources such as podcasts or audiobooks. To manually process sufficient data – for example, transcribing the words – would be time consuming and expensive, and thus a barrier to creating speech recognition or synthesis systems for new domains or new languages.

Automatic alignment of speech with imperfect transcripts has already been well addressed in the previous work of others, for example [1, 2, 3, 4, 5, 6, 7]. Unfortunately, all of these approaches make use of expert knowledge and/or expensive resources, such as very good speaker-independent acoustic models or large vocabulary 'biased' language models, and therefore can only be applied to languages where these resources exist.

In our own previous work [8, 9, 10], we introduced a lightly supervised method for automatically aligning speech data with imperfect transcripts that does not rely on such resources. Our method comprises two main components: a GMM-based sentence-level segmentation algorithm, and an alignment step which uses incrementally-trained grapheme-level acoustic models to determine the correct orthographic transcription of the segmented utterances. Both steps are lightly supervised, in the sense that they need only small amounts of manual initialisation before proceeding in a fully automatic way with no further intervention from the user, and all statistical models used are learned solely from the speech and text being aligned. Baseline results and evaluations were obtained using a Librivox audiobook recording[1] of *A Tramp Abroad* by Mark Twain, but we have since successfully applied the algorithms to audiobooks in 14 different languages, thus creating the TUNDRA corpus [11].

The success of a speech/text alignment algorithm can be quantified in terms of amount of speech data 'harvested' with correctly aligned transcriptions. While building the TUNDRA corpus, we found that most of the Librivox audiobooks we used had recording conditions that were highly variable within a single book across the different chapters, and that this led to lower harvesting rates. That is, a lower percentage of the data was aligned than expected, especially for chapters where the recording conditions were very different from the book average. Although it can be argued that this noisy data would be better left out of the final speech resource, in many applications the amount of training data is more important than its recording quality and maximising the amount of data aligned is the primary concern. We have therefore been investigating ways to improve the amount of data harvested.

In this paper we apply two adaptation methods to the two main stages of our method: Maximum A Posteriori (MAP) adaptation for the GMM-based segmentation algorithm, and Constrained Maximum Likelihood Linear Regression (CMLLR) transform-based adaptation for the acoustic models (HMMs) used in the alignment step. We show that by employing these techniques, our alignment results for noisy data significantly improve in both the percentage of data aligned

[1] http://librivox.org/a-tramp-abroad-by-mark-twain/ read by John Greenman

and in the accuracy of the aligned data. Although these are standard adaptation procedures, there were some challenges in using them in this context: for MAP, we need to devise a process for selecting the adaptation data in accordance with the specific structure of audiobooks; for CMLLR, the lack of accurate transcripts for the adaptation data, and the use of grapheme-level acoustic models, posed particular problems.

The paper is structured as follows: Sections 2 and 3 present the adaptation methods used for the segmentation and alignment stages respectively. The results obtained with these methods on sets of noisy data are evaluated in Section 4, while Section 5 concludes the paper and discusses future work.

2. Lightly Supervised Speech Segmentation using MAP Adaptation

In [8] we proposed a lightly supervised sentence-level segmentation tool based on Gaussian Mixture Models (GMM) which is an extension of a method widely used for Voice Activity Detection (VAD). The core idea was to train two GMMs: one from the speech segments, and the other from the silence segments, of an initial manually-labelled data set of only 10 minutes of speech. The GMMs were then used to estimate the log likelihood of all segments of the full data being silence or speech. Because short silent pauses can occur within running speech, the algorithm was tuned to detect only sentence boundaries, and not within-sentence pauses. A threshold for discriminating between short pauses and silence was automatically calculated by fitting two Gaussians (one for extended silence and one for short pauses) to the durations of these two types of silent sections, using the manually-labelled data. Results showed a 96% accurate detection rate. Another aspect of performance that we evaluated was the effect of this VAD-based segmentation on the final quality of synthetic voices built from this data. By training two text-to-speech systems, one with a GOLD standard (i.e., manually verified and corrected) segmentation and one with the VAD-based one, we determined that the VAD-based voice had only a marginally, and statistically insignificantly, lower quality.

While the above techniques work well on consistent, clean speech, when used on the data being prepared for the TUNDRA corpus it was found that using GMMs trained only on the small set of manually-labelled data did not give good performance across all the remaining data. This was because they did not capture the correct distributions for silence and speech in the varying noise environments and speaking styles. Therefore, we propose a method to adapt the initial GMMs on a chapter-by-chapter basis. The workflow employed in performing this adaptation and segmentation is presented in Figure 1 and comprises the following steps:

1. **Initial training** – initial GMMs for speech and silence are trained on the labelled data;
2. **1st decoding** – label the speech and silence parts of all chapters using these initial GMMs;
3. **Data selection** – apply a confidence measure to each such speech or silence part, selecting only the *confident* data;
4. **MAP adaptation** – adapt the GMMs using a standard MAP algorithm [12, 13, 14, 15] to this data;
5. **2nd decoding** - re-label the speech and silence parts of all chapters using the adapted GMMs. Segment the chapter at every silence mid-point.

The data selection step described above is used to select the speech and silence segments which are considered to be confidently-labelled and thus suitable as adaptation data. The confidence score or measure we use is based on a log likelihood ratio (LLR) computed for each segment against the respective GMM. Figure 2 shows an example of these histograms for the speech and silence parts of the data corrupted with *babble talk noise at 35dB*: (a) represents the LLR histogram without adaptation; (b) is the histogram after performing MAP adaptation on all the speech and silence parts resulting from the 1st decoding; and (c) is the histogram after MAP adaptation using only the confident segments. This shows that performing adaptation using all the data from the 1st decoding leads to mis-classification of audio segments: the discriminative power of the GMMs is reduced by 18% (compare distance between peaks in Figure 2 (a) with (b)). On the other hand, if data selection is carried out, the histogram plots show an increase in the distance between the average LLRs by 11% (Figure 2 (b) with (c)).

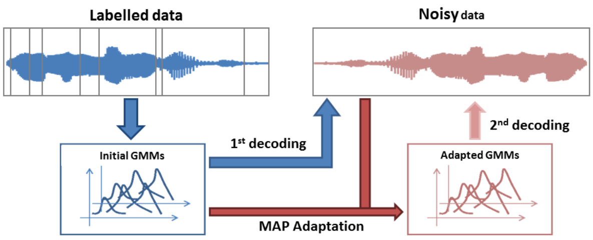


Figure 1: Overview of the MAP adaptation method for the GMM-based VAD.

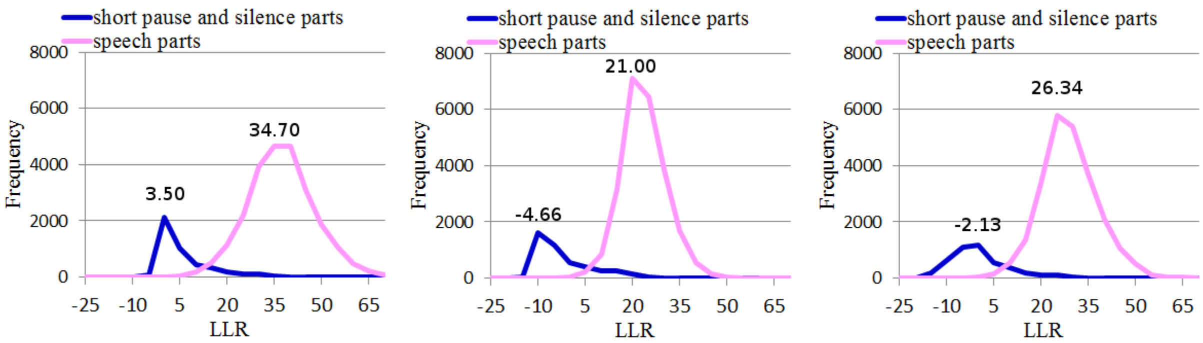
3. Speech Transcription Alignment using CMLLR Adaptation

[9] introduced a lightly supervised and low-resource method for sentence-level alignment of speech with imperfect transcripts. The method incrementally trains grapheme-level acoustic models on the available speech and text data, starting from an initial 10 minute manual orthographic speech transcription[2]. In order to alleviate the consequences of having rather poor acoustic models, the Viterbi decoder was highly restricted by using a so-called *skip network*. The network allows the speech to be matched to any point within an estimated broad text window, but constrains the output to be a consecutive sequence of words from it. To deal with audio deletions, a more relaxed skip network, called a *3-skip* network, can be used which allows a maximum 2 word skip within the hypothesised sequence. To prevent unwanted skips, a bigram language model built from the available text was also used to limit to some extent the 3-skip network. The confidently aligned utterances were then obtained by comparing the recognition acoustic scores using the different types of skip networks. These utterances were then used to retrain the acoustic models, and the process repeated for a couple of iterations. Results from this method showed a 54.1% aligned percentage with 7.64% SER (sentence error rate) and 0.5% WER (word error rate).

Following this, in [10] we increased the alignment percentage by almost 40% (relative) through the use of context-dependent tri-grapheme models and MMI discriminative training. The confidently aligned data then amounted to 75%, with similar sentence and word error rates as in the previous work.

Despite this good performance on our test audiobook (for which we have the GOLD standard alignments required to com-

[2] The same 10 minutes of labelled speech data for the VAD can be used for the aligner as well



(a) VAD without MAP adaptation

(b) MAP VAD without data selection

(c) MAP VAD with data selection

Figure 2: Segment LLR histograms for silence and speech data calculated for (a) VAD without MAP adaptation, (b) MAP VAD without data selection and (c) MAP VAD with data selection. The data on which they are estimated was corrupted with *babble talk noise at 35dB*.

pute SER and WER), when applying the above procedures to other audiobooks from Librivox, we found that the variable recording conditions across chapters (e.g. more background noise as a result of variable distance from the microphone, or worse room acoustics) caused the aligned percentage to drop below 40% for the worst chapters.

To address this problem, we turn to adaptive training methods commonly used in automatic speech recognition and in this paper we propose the use of CMLLR [16], originally proposed for speaker adaptation, for adaptation to the varying channel or environmental conditions found in audiobooks. The CMLLR technique estimates a set of linear transforms for each condition—shared between multiple Gaussians—in a maximum likelihood fashion, making it robust to estimation when the initial transcripts are poor, and allowing effective use of limited adaptation data.

Here we apply CMLLR adaptation to the discriminatively-trained tri-grapheme acoustic models presented in [10] to adapt to the noisy data. Although the poor recognition accuracy over the noisy speech means that the quality of the adaptation transcripts is also quite low, the results show that by using only one or two chapters as adaptation data, the SER and the percentage of aligned data are substantially improved.

4. Results

4.1. Simulated Noisy Speech Recordings

To test our approach, it is necessary to be able to compute SER and WER, for which we need GOLD transcripts. We therefore once again used the Librivox audiobook *A Tramp Abroad* by Mark Twain and degraded two chapters of the audiobook (approx. 28 minutes of speech) by adding noise and/or reverberation to simulate the noisy data found in real recordings in a controllable way.

Through informal evaluation, we determined which conditions approximated those observed in the TUNDRA corpus and similar found data. For reverberation, we convolved the speech with the impulse response of a domestic living room, taken from the Open Air Library [17]. The background noise conditions were replicated using either 8-talker babble or white noise at the following signal-to-noise ratios (SNRs): 10, 15, 20, 25, 30, 35 and 40dB. A total of 29 testing scenarios were obtained this way: reverb; babble noise at each SNR; white noise at each SNR; reverberation and babble noise at each SNR; reverberation and white noise at each SNR. Although 10dB and 15dB SNRs are highly unlikely (i.e., very noisy) for audiobook recordings, these scenarios were kept as points of comparison to evaluate the adaptive power of the acoustic models even when the accuracy of the transcripts is very low.

4.2. GMM VAD with MAP adaptation

We present the evaluation of three versions of VAD: without MAP adaptation; with MAP adaptation, but without data selection; and with MAP adaptation and with data selection. The CORR measure is computed as [18]:

$$CORR = 100 - (FEC + MSC + OVER + NDS) : \quad (1)$$

where the right hand side measures represent (as percentages):

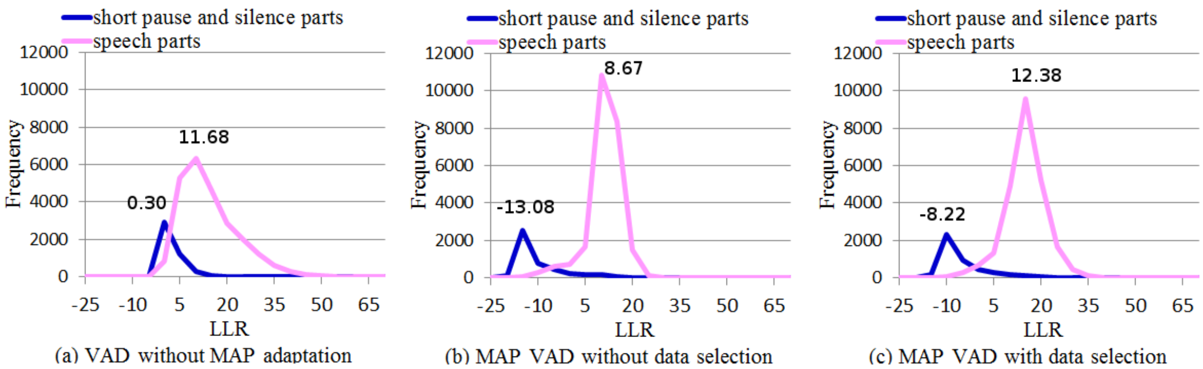
- **FEC** - Front End Clipping - speech classified as silence when passing from silence to speech;
- **MSC** - Mid Speech Clipping - speech classified as silence within a speech sequence;
- **OVER** - silence interpreted as speech at the end of a speech segment;
- **NDS** - Noise Detected as Speech - silence interpreted as speech within a silence part.

Figure 4 shows the CORR measure for each environment. The results show a high dependency on the type of noise and the SNR. For white noise, MAP adaptation gave great improvements at high SNRs. At low SNRs, because of the fact that the initial GMMs were unable to discriminate between speech and silence—all segments were labelled as speech—there are no differences in the CORR measure for the 3 VAD types. The high value of the CORR measure is a result of the fact that the speech segments are much longer than the silence segments, and this influences the FEC and MSC values.

For babble noise, there are noticeable advantages of using MAP, but only for mid-range SNR values. At low and high SNRs, the CORR value is similar to that when no adaptation is used.

When adding reverberation to the clean data, MAP adaptation performs better without the data selection step. This is also true in the case of reverberation and babble talk noise. This may be due to mismatch of the threshold for data selection in these environments. We used the threshold which was the most appropriate for 35dB of babble noise across all environments.

In contrast, VAD without MAP adaptation showed higher CORR across all SNRs for reverberation plus white noise. But



(a) VAD without MAP adaptation (b) MAP VAD without data selection (c) MAP VAD with data selection

Figure 3: Segment LLR histograms for silence and speech data calculated for (a) VAD without MAP adaptation, (b) MAP VAD without data selection and (c) MAP VAD with data selection. The data on which these examples were estimated was corrupted with *reverberation and white noise at 15dB.*

when examining the LLR histogram for this condition (see Figure 3), adaptation seems increase the discriminative power of the GMMs. This leads us to believe that white noise plus reverberation has the most damaging effect on the GMM-based VAD, and that alternative methods for dealing with this type of scenario must be investigated.

4.3. CMLLR Acoustic Model Adaptation

As described in Section 3, the baseline acoustic models built on the clean data were adapted using the simulated noisy speech data. The adaptation transcripts were obtained from the baseline models using a 1-skip network. For each noisy scenario, we computed the SER and WER of the entire noisy data, as well as the amount of confident data obtained (i.e., the percentage aligned) with its corresponding SER. WERs for the confident data are on average below 1%, do not seem to be influenced by the adaptation step, and are therefore not presented.

Figures 5 and 6 present the SER and WER of the entire noisy data respectively. The SER and WER values are computed for the text aligned using the adapted acoustic models, as compared to the GOLD standard transcripts. As expected, SER and WER are reduced by the use of adaptation, especially at low SNRs. The type of noise has a strong influence on the overall performance: white noise in conjunction with reverberation has the most damaging effect on the performance of the clean acoustic models. One other thing worth noting is the fact that, although the SER in some cases is quite high, the corresponding low WER makes adaptation possible. For example in the case of white noise at 20dB SNR, the SER is around 70%, but the WER is around 20%. The adaptation in this case improves both the WER and SER of the entire noisy data by a substantial amount (approx. 50% for SER and 20% for WER).

The improvement in the SER and WER of the noisy data through adaptation would mean nothing unless it also influences the aligned data percentage. In figure 7 we present this influence. The bars in the figure represent the percent of confident data with its relative SER. Again, the adaptation makes the percent of confident data increase, and also lowers its SER. The average increase in confident data percentage is 20%, with a maximum of 62% for the reverberation and white noise at 25dB scenario. In extreme cases, the adaptation did not help (such as white noise at 10dB and 15dB, reverberation and babble noise 10dB, reverberation and white noise 10dB and 15dB), but these are almost certainly not of interest anyway, if the speech is going to be used to build a speech synthesiser.

Note that the numbers presented in this section are not directly comparable to those in [10], because here we are evaluating only a small subset of the audiobook, and not its entirety.

5. Conclusions

In this paper we have shown the advantages of using adaptive techniques in order to improve the alignment accuracy of text with corresponding noisy and/or reverberated speech, which for experimental purposes was created by simulating the conditions we have observed in various typical non-professional audiobooks.

The speech segmentation algorithm performance is highly dependent on the noise characteristics, giving variable improvements across the tested scenarios. The presence of reverberation leads to unexpected results in terms of the CORR measure, and the white noise plus reverberation renders the adaptation ineffective. On the other hand, when applying adaptation to the acoustic models of the speech aligner, the amount of confident data increases in all scenarios, resulting in an average 20% improvement. It also reduces the SER of this data.

Future work includes the evaluation of the effect of adaptation for both segmentation and alignment on the confident data percentage. Another technique which can be employed is to cluster data based on recording environments, and do cluster-based adaptation (rather than chapter-based). We would also like to investigate the influence of the VAD indices (CORR, FEC, MSC, OVER and NDS) on the TTS system's quality when using various noise environments and amount of adaptation data.

6. Acknowledgements

The GOLD transcripts for *A Tramp Abroad* were very kindly provided by Toshiba Research Europe Limited, Cambridge Research Laboratory. The research leading to these results was partly funded from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement 287678 (Simple4All), EPSRC EP/I031022/1 (NST) and EP/J002526/1 (CAF). The research presented here has made use of the resources provided by the Edinburgh Compute and Data Facility (ECDF: `http://www.ecdf.ed.ac.uk`). The ECDF is partially supported by the eDIKT initiative (`http://www.edikt.org.uk`).

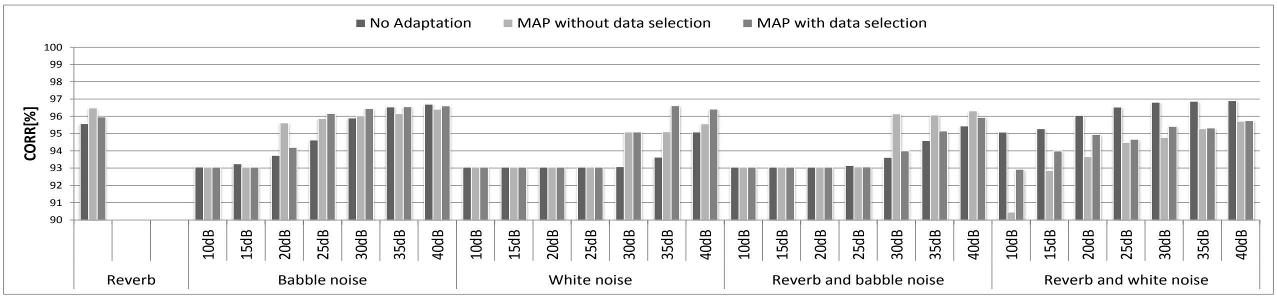


Figure 4: CORR measure for each VAD method: using no adaptation, with MAP adaptation but without data selection and with MAP adaptation and data selection.

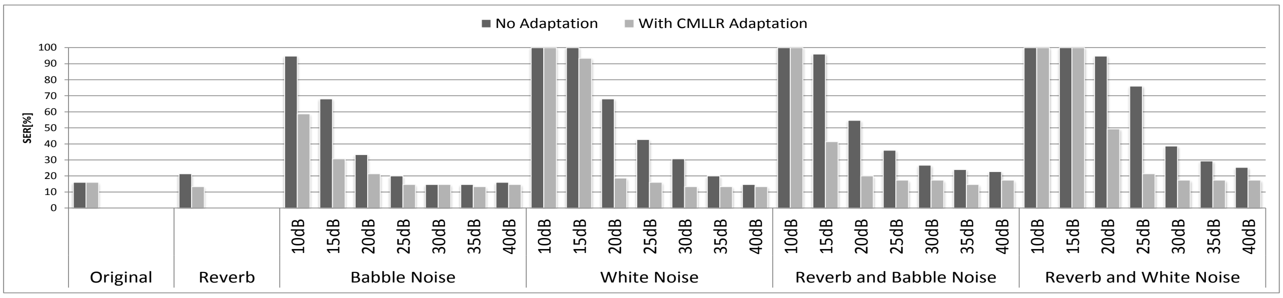


Figure 5: SER for each noisy and reverberant data set, with and without CMLLR adaptation. SER is computed on the retrieved text for each acoustic model against a GOLD standard transcription.

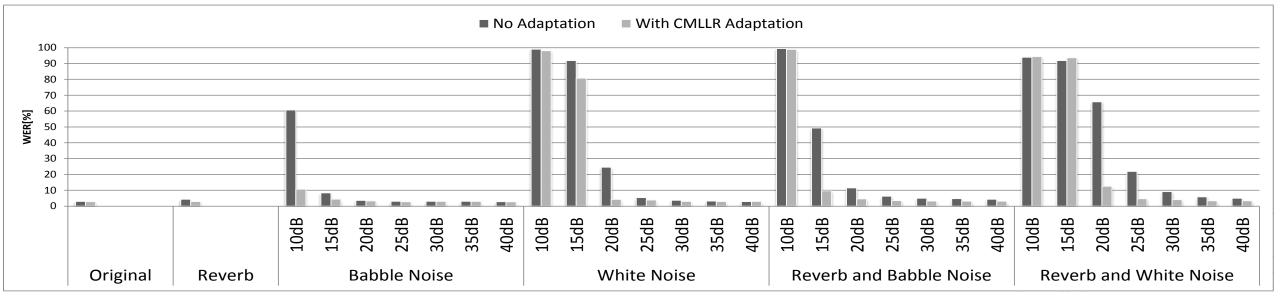


Figure 6: WER for each noisy and reverberant data set, with and without CMLLR adaptation. WER is computed on the retrieved text for each acoustic model against a GOLD standard transcription.

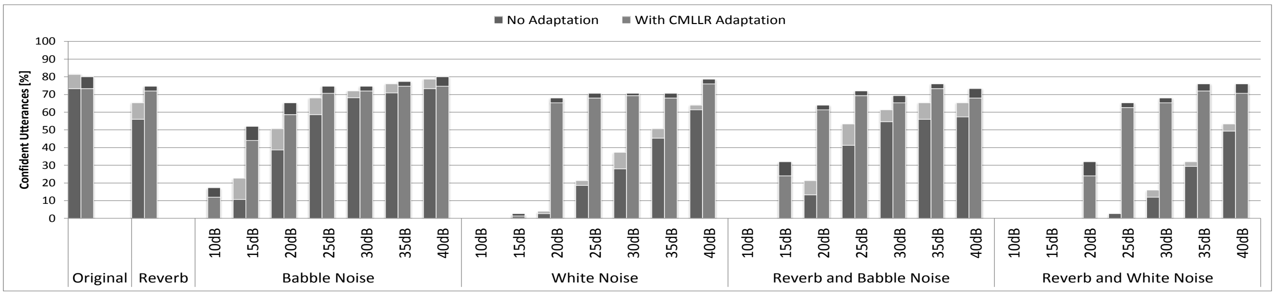


Figure 7: Aligned data percentage obtained before and after CMLLR adaptation. The different colour bar at the top of each column represents the relative SER for each data set.

7. References

[1] É. Székely, T. G. Csapó, B. Tóth, P. Mihajlik, and J. Carson-Berndsen, “Synthesizing expressive speech from amateur audio-book recordings,” in *Proc. IEEE Workshop on Spoken Language Technology*, Miami, Florida, USA, 2012, pp. 297–302.

[2] O. Boeffard, L. Charonnat, S. L. Maguer, and D. Lolive, “Towards Fully Automatic Annotation of Audio Books for TTS,” in *Proc. of LREC*, Istanbul, Turkey, may 2012.

[3] N. Braunschweiler, M. Gales, and S. Buchholz, “Lightly supervised recognition for automatic alignment of large coherent speech recordings,” in *Proc. of Interspeech*, 2010, pp. 2222–2225.

[4] K. Prahallad, A. R. Toth, and A. W. Black, “Automatic building of synthetic voices from large multi-paragraph speech databases,” in *Proc. of Interspeech*, 2007, pp. 2901–2904.

[5] P. Moreno and C. Alberti, “A factor automaton approach for the forced alignment of long speech recordings,” in *Proc. of ICASSP*, 2009, pp. 4869–4872.

[6] G. Bordel, M. Peñagarikano, L. J. Rodríguez-Fuentes, and A. Varona, “A simple and efficient method to align very long speech signals to acoustically imperfect transcriptions,” in *Proc. of Interspeech*, 2012.

[7] M. Alessandrini, G. Biagetti, A. Curzi, and C. Turchetti, “Semi-Automatic Acoustic Model Generation from Large Unsynchronized Audio and Text Chunks,” in *Proc. of Interspeech*, 2011, pp. 1681–1684.

[8] Y. Mamiya, J. Yamagishi, O. Watts, R. A. J. Clark, S. King, and A. Stan, “Lightly Supervised GMM VAD to use Audiobook for Speech Synthesiser,” in *Proc. ICASSP*, 2013.

[9] A. Stan, P. Bell, and S. King, “A grapheme-based method for automatic alignment of speech and text data,” in *Proc. IEEE Workshop on Spoken Language Technology, Miami, Florida, USA*, 2012.

[10] A. Stan, P. Bell, J. Yamagishi, and S. King, “Lightly Supervised Discriminative Training of Grapheme Models for Improved Sentence-level Alignment of Speech and Text Data,” in *Proc. of Interspeech (accepted)*, 2013.

[11] A. Stan, O. Watts, Y. Mamiya, M. Giurgiu, R. A. J. Clark, J. Yamagishi, and S. King, “TUNDRA: A Multilingual Corpus of Found Data for TTS Research Created with Light Supervision,” in *Proc. of Interspeech (accepted)*, 2013.

[12] Y. Zhang and M. S. Scordilis, “Effective online unsupervised adaptation of gaussian mixture models and its application to speech classification,” *Pattern Recognition Letters*, vol. 29, no. 6, pp. 735–744, 2008.

[13] J. Gauvain and C. Lee, “Maximum a posteriori estimation for multivariate gaussian mixture observations of Markov chains,” *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 2, pp. 291–298, 1994.

[14] T. Oonishi, K. Iwano, and S. Furui, “Noise-robust speech recognition decorder using speech/non-speech confidence measures,” *Technical Report of IEICE*, vol. 110, no. 81, pp. 49–54, 2010.

[15] D. A. Reynolds, T. F. Qatieri, and R. B. Dunn, “Speaker verification using adapted gaussian mixture models,” *Digital Signal Processing*, vol. 10, no. 1, pp. 19–41, 2000.

[16] M. Gales, “Maximum likelihood linear transformations for HMM-based speech recognition,” *Computer Speech and Language*, vol. 12, pp. 75–98, 1998.

[17] S. Shelley, A. Foteinou, and D. Murphy, “OpenAIR: An Online Auralization Resource with Applications for Game Audio Development,” in *Proceedings of the 41st Int. Conference of the AES, Audio for Games*, 2011.

[18] F. Beritelli, S. Casale, and G. Ruggeri, “Performance evaluation and comparison of ITU-T/ETSI voice activity detectors,” in *Proc. ICASSP*, vol. 3, Salt Lake City, UT, USA, 2001, pp. 1425–1428.