Lightly Supervised Discriminative Training of Grapheme Models for Improved Sentence-level Alignment of Speech and Text Data

Adriana Stan[1], *Peter Bell*[2], *Junichi Yamagishi*[2,3], *Simon King*[2]

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

adriana.stan@com.utcluj.ro, {peter.bell, jyamagis, simon.king}@inf.ed.ac.uk

Abstract

This paper introduces a method for lightly supervised discriminative training using MMI to improve the alignment of speech and text data for use in training HMM-based TTS systems for low-resource languages. In TTS applications, due to the use of long-span contexts, it is important to select training utterances which have wholly correct transcriptions. In a low-resource setting, when using poorly trained grapheme models, we show that the use of MMI discriminative training at the grapheme level enables us to increase the amount of correctly aligned data by 40%, while maintaining a 7% sentence error rate and 0.8% word error rate. We present the procedure for lightly supervised discriminative training with regard to the objective of minimising sentence error rate.

Index Terms: automatic alignment, grapheme models, light supervision, discriminative training

1. Introduction

Recent advances in HMM-based text-to-speech synthesis (TTS) have made the use of large volumes of imperfect, natural speech data an essential aspect in developing systems for new speakers, dialects or even whole languages. When sourcing expressive speech data for system building, we have the choice of recording the speaker under a carefully controlled conditions, or else to manually transcribe and annotate several hours of an existing speech corpus. Neither option is efficient for developers seeking to control the selection of multiple speakers or speaking styles for building synthetic voices.

Another disadvantage of the conventional speech databases is the fact that they are in general recorded as individual utterances, with no correlation in between them or across a paragraph or recording session.

To alleviate this problem, the research community has recently shifted focus towards the use of audiobooks as a readily available, more expressive speech resource [1, 2, 3, 4, 5]. However, the use of audiobook data, not specifically recorded for speech technology applications, requires a reliable matching transcription to be obtained. The use of an existing ASR system for this purpose has been proposed by many researchers [3, 6, 7, 8, 9, 10, 11]. However, these methods are applicable only to languages where the resources for training a good speaker-independent ASR system already exist. For an under-resourced language, the only audio data available may be audiobook data for the target speaker. In this case, a method is required which is able to use only this data, with possibly unreliable transcriptions, with no bootstrapping from other data.

The approach taken in this study follows from our previous work [12] where we used lightly supervised acoustic model training, somewhat similar to that in [13], using only grapheme models instead of phone models, under the assumption that a manually-created phone dictionary may not be available for an arbitrary new language. This makes the alignment problem harder, since grapheme-based acoustic models (AM) perform worse than phone-based ones [14]. In our earlier work, the success of the method relies on us making a conservative selection of data where the automatically-retrieved transcription is judged to be reliable, using the poor acoustic models.

This paper investigates methods for improving the performance by the use of discriminative training of the context-dependent grapheme models. Although the use of discriminative training – using, for example, the MMI or MPE criterion – is widespread in ASR systems, it is not common in TTS applications, where the generative nature of the HMMs is important. However, methods such as MMI are known to improve on ML training where the underpinning model correctness assumption does not hold [15], a particular problem for grapheme-based models. Furthermore, as we discuss in 3.1 the MMI criterion is better matched to the objective of minimising SER – this is important, because in TTS applications, due to the role of long-span contexts in model training, it is more important to obtain utterances which have wholly correct transcriptions.

2. Low-resource alignment of speech and text data

In our previous study [12], we presented an unsupervised, language independent method for aligning speech data with imperfect transcripts. The method relies solely on the available speech and text resources, and uses a highly restricted word network to perform a Viterbi alignment with the poor grapheme-level acoustic models for each segment. This word network, which we term a *skip network* is built using the available text. The core principle is to allow the alignment to begin or end at any point in the text, whilst constraining the word ordering to match word strings in the text, up to a maximum of 2 word deletions or skips. Further, network paths not found in a bigram language model derived from the same text are removed (Fig. 1 illustrates the network). This network imposes much tighter constraints on the output than the traditional biased language model approaches such as [7]. Using the skip network, we were able to extract 54.1% of utterances with high confidence, with the scores of 7.64 % SER and 0.5 % WER on the recovered text. A confidence measure for the aligned data is obtained by comparing the acoustic scores of the recognised output, using

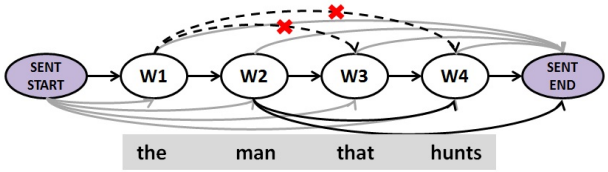

Figure 1: Word skip network design.

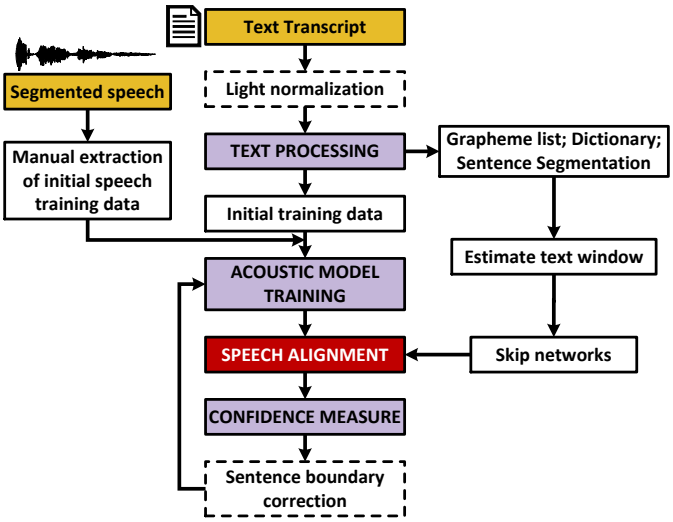

Figure 2: Flowchart of the training and alignment process.

different degrees of freedom in the skip network. Some light pre- and post-processing of the data is performed, as presented in Fig. 2.

3. Acoustic model training

3.1. Discriminative objective functions for SER minimisation

In conventional maximum likelihood training for HMM-based TTS systems, we aim to find parameters θ which best explain the training data, in the sense of maximising, over all training utterances, r, the joint likelihood of acoustic observations O_r and word sequences W_r. This leads to the objective function

$$F_{\text{ML}}(\theta) = \sum_r \log p_\theta(O_r|W_r)P(W_r) := \sum_r D_\theta(O_r, W_r) \tag{1}$$

where $D_\theta(O, W) = \log p_\theta(O|W)P(W)$ is a discriminant function. In contrast, discriminative training seeks to explicitly consider the classification decisions made by the resulting model. If the objective is to maximise the number of completely correct sentences, minimising SER, a natural choice of objective function is the Minimum Classification Error (MCE) criterion [16], considering the error on a complete-sentence basis. This function is, however, difficult to optimise. An alternative is to maximise the margin $\mathcal{E}_r$ by which utterances are correctly classified,

$$\mathcal{E}_r = D_\theta(O_r, W_r) - \max_W D_\theta(O_r, W) \tag{2}$$

which leads to the objective function

$$F(\theta) = \sum_r \mathcal{E}_r \tag{3}$$

$$\geq \sum_r \Big[D_\theta(O_r, W_r) - \log \sum_W e^{D_\theta(O_r, W)} \Big] \tag{4}$$

$$= \sum_r \log \frac{p_\theta(O_r|W_r)P(W_r)}{\sum_W p_\theta(O_r|W)P(W)} = F_{\text{MMI}}(\theta) \tag{5}$$

where we use the softmax approximation to derive a lower bound that is easier to optimise. This is the well-known MMI criterion for discriminative training. It can be shown [15, 17] that the expected error rate using MMI-trained models converges to the model-free expected error rate as the amount of training data increases: in other words, MMI does not require the correct generative model to be used in order to be effective. In practice, the acoustic probabilities in Equation 5 are scaled by the inverse of the language model scaling factor; the sum over all words W in the denominator is computed over a lattice.

3.2. Lightly supervised MMI training

For the practical application of MMI training in the low-resource, lightly supervised setting, a number of issues must be considered. Firstly, since the MMI objective function is based on the difference between numerator and denominator terms for each utterance r, it is important to select only utterance for which we already have a confident text alignment. This problem has been considered by [18]. Secondly, we must consider how the the denominator lattices should be constructed. In ASR applications, a weaken version of the language model used for recognition would be used: for TTS, there is no such model, and for a new language, we may not have good coverage to build one. An alternative is to generated denominator lattices over graphemes, similar to early approaches for discriminative training for ASR, such as [19], where phone-level lattices were used.

In this work, the confident utterances were selected using the alignment and confidence measure described in [12] and for which we showed a WER of less than 1%. For denominator lattices instead, we compared the use of both word level and grapheme-level lattices. Both word and grapheme level training used a bigram language model derived from the original text. But for the grapheme level lattices, the text was first converted into grapheme sequences. Details of the full implementation are given in Section 3.3.

3.3. Training procedure

To test our hypothesis, we built 4 different types of acoustic models starting from the five-state, left-to-right, mono-graphemes with eight mixture components per state, and no state tying.

The first step in building the final acoustic models was to extend the mono-grapheme models to tri-grapheme models, using a standard procedure. The list of tri-graphemes was extracted from the text, so the models will not generalise well to unseen text but are adequate for aligning the available text, though to reduce over-fitting, we model only within-word context. This reduces the number of models. Eight re-estimations were performed, using the confident transcripts of the mono-grapheme ML trained models.

As discussed in Section 3.2, for discriminative training two approaches were selected, one using word-level lattices, the

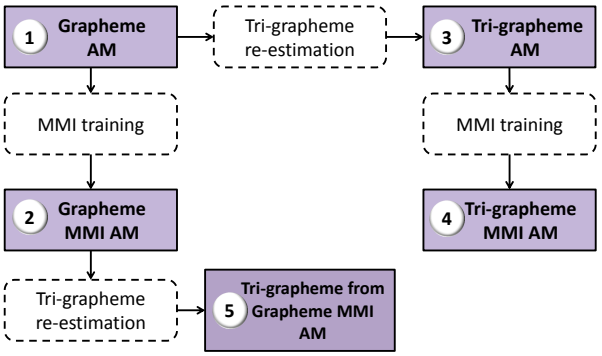

Figure 3: *Schematic diagram of the acoustic model training method.*

other using grapheme-level ones. In the word-level training, we only used the tri-grapheme model. Denominator word lattices were built by running the tri-grapheme models with a bigram LM derived from the original text, over the training data. Numerator lattices were generated from the approximate transcripts obtained in the alignment process. For the grapheme-level training, the entire text was converted into grapheme sequences and a grapheme language model was built. Numerator and denominator lattices were obtained in a similar way to that presented above.

The generated lattices and the approximate transcriptions were used to find the grapheme model boundaries and produce denominator and numerator model-marked lattices. The MMI models were obtained by re-estimating the ML models over 8 iterations, with a grammar scale factor of 30. The schematic diagram of the entire AM training process is presented in Fig.3.

4. Results

In order to test the lightly supervised discriminative training method, we used an audiobook of *A Tramp Abroad* by Mark Twain[1], for which a gold-standard transcription was kindly provided by Toshiba Reseach Europe Limited, Cambridge Research Laboratory. All SER and WER figures are computed with reference to this transcription.

We first investigated the effect of using poor initial alignments on the effectiveness of MMI training, as well as the quantity of training data. To do this, we applied MMI training to to initial ML-trained mono-grapheme models using different versions of the reference transcription. Table 1 presents result with four cases: `GOLD` denotes the entire audiobook data with the ground-truth transcriptions; `CONF` is only the confident utterances selected by the alignment procedure using the first iteration acoustic model (approx. 54% of the data); `ALL` is the entire data with transcriptions obtained using the first iteration acoustic model; finally, `IT-0` represents the data obtained using the very initial acoustic model, which was trained on only 10 minutes of data. From the results, it may be observed that the MMI training is relatively robust to the use of possibly incorrect transcriptions. Even in the case of `IT-0`, we find that the use of MMI training results in a SER reduction of 2.5% and a WER reduction of 0.5%, compared to figures reported in our previous study of 22% and 3% respectively

One other evaluation refers to the use of the two different lattice building methods, word and grapheme level lattices. 5 acoustic models are evaluated from a SER and WER point-of-view, and the results are presented in Table 2. The numerator lattices are built from the confident data obtained with the baseline AM. Acoustic model descriptions are as follows[2]: `(1)ML-MG` - baseline ML trained mono-grapheme acoustic models; `(2)ML-TG` - ML trained tri-grapheme acoustic models; `(3)MMI-MG` - MMI trained mono-grapheme AM; `(4)MMI-TG` - MMI trained tri-grapheme AM with word-level lattices; `(5)TG-MMI-MG` - MMI trained mono-grapheme AM with grapheme-level lattices and then re-estimated into tri-grapheme models. The overall improvement going from baseline mono-grapheme (`ML-MG`) models to tri-grapheme MMI (`TG-MMI-MG`) is of 6.5% in SER and 0.6% in WER. This means that even in the given poor acoustic models, this method manages to correctly align almost 90% of the original data. However, the final results are also influenced by the confidence measure, which has its errors, also.

Table 1: *The influence of the amount and quality of training data transcription over the accuracy of the grapheme MMI model.*

Training Data	SER [%]	WER [%]
GOLD	17.62	2.29
CONF	17.40	2.24
ALL	17.80	2.26
IT0	19.50	2.49

Table 2: *SER and WER for the entire data using different acoustic models and the confident utterances as training data*

Model	SER [%]	WER[%]
(1) ML-MG	18.72	2.43
(2) ML-TG	17.44	5.88
(3) MMI-MG	17.40	2.24
(4) MMI-TG	14.98	3.59
(5) TG-MMI-MG	12.15	1.84

Note that the most important goal of the whole process is to be able to extract greater quantities of data from the speech and text resources available. Considering this aim, Table 3 presents the proportion of confident alignments obtained by each of the above acoustic models, along with their SER and WER. The acoustic model description is as above. A first thing to observe is that fact that simply by moving from the ML mono-grapheme models to ML tri-grapheme models, the number of confident alignments increases by 30%, going from 54.2% to 70.1% while maintaining the approximately the same error rates. The best performing model, as expected is the tri-grapheme dervived from the MMI-trained mono-graphemes which add 40% confident data for the same levels of error. These results are comparable in WER to those presented in [7].

5. Discussion

Unsupervised or lightly supervised speech and text alignment is an important prerequisite in the process of building text-to-speech systems in a language or domain where language analysis and acoustic model building are tedious (i.e. annotated speech resources are scarce and there is little or no language ex-

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

[2] Numbers correspond to those in Fig. 3

Table 3: *The percentage of confident data and its SER and WER for each acoustic model*

Model	Percentage [%]	SER [%]	WER [%]
(1) ML-MG	54.10	7.64	0.5
(2) ML-TG	70.11	8.29	1.32
(3) MMI-MG	59.29	9.85	0.57
(4) MMI-TG	70.37	7.54	0.87
(5) TG-MMI-MG	75.88	7.59	0.80

pertise). The method presented here, though, showed that using only speech data and its approximate orthographic transcript, almost 70% can be aligned[3]. The improvements obtained by both the tri-grapheme and lightly supervised acoustic models are significant, and amount to an overall increase of 40% while maintaining the word and sentence error rates. In a separate study, we obtained similar numbers for other 14 languages. Also, [20] showed no significant difference between TTS voices built on manually aligned data, and our ML approach. Therefore by increasing the amount of confident data using tri-graphemes and discriminative training, the synthesis quality can only improve.

But still, the major goal of this ongoing work is to be able to align the entire speech resource available by taking into account audio insertions and substitutions as well. Confidence measure can also be improved so that the error rates of the confident files are next to 0%. Also, from the difference between the computed accuracy of the models, and the percentage of confident files emphasizes the need for a better confidence measure. One other aspect which might occur in found data is the different recording conditions across sections or chapters of the speech resource. A way to overcome the potential alignment problems would be to use multi-condition or speaker adaptive training.

6. Conclusions

This paper introduced a series of enhancements for poor grapheme acoustic models used in the alignment of speech and text data. Extending the mono-grapheme models to tri-grapheme ones and then performing discriminative training at word and grapheme level increased the percent of extracted confident data by 40% in the same SER and WER conditions. Compared to other studies on limited domain recognition, our methods are based only on the available data and use minimal user intervention or language expertise.

7. Acknowledgements

The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement no. 287678. (Simple4All). The GOLD transcripts for *A Tramp Abroad* were very kindly provided by Toshiba Research Europe Limited, Cambridge Research Laboratory. The grapheme-based acoustic model training tools were provided by Oliver Watts.

[3]Confident percentage is 75%, but taking into account the 7% SER, we are left with 70% fully correct data

8. References

[1] L. Chen, M. J. F. Gales, V. Wan, J. Latorre, and M. Akamine, "Exploring Rich Expressive Information from Audiobook Data Using Cluster Adaptive Training," in *Proc. of Interspeech*, 2012.

[2] É. Székely, J. P. Cabral, M. Abou-Zleikha, P. Cahill, and J. Carson-Berndsen, "Evaluating expressive speech synthesis from audiobook corpora for conversational phrases," in *Proc. of LREC*, 2012, pp. 3335–3339.

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

[4] K. Prahallad and A. W. Black, "Segmentation of Monologues in Audio Books for Building Synthetic Voices," *IEEE Transactions on Audio, Speech & Language Processing*, vol. 19, no. 5, pp. 1444–1449, 2011.

[5] N. Braunschweiler and S. Buchholz, "Automatic Sentence Selection from Speech Corpora Including Diverse Speech for Improved HMM-TTS Synthesis Quality," in *Proc. of Interspeech*, 2011, pp. 1821–1824.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] A. Stan, P. Bell, and S. King, "A Grapheme-based Method for Automatic Alignment of Speech and Text Data," in *Proc. of IEEE Workshop on Spoken Language Technology*, Miami, Florida, USA, Dec. 2012, pp. 286–290.

[13] L. Lamel, J. luc Gauvain, and G. Adda, "Lightly Supervised and Unsupervised Acoustic Model Training," *Computer Speech and Language*, vol. 16, pp. 115–129, 2002.

[14] M. Killer, S. Stüker, and T. Schultz, "Grapheme based speech recognition," in *Proc. of Interspeech*, 2003.

[15] R. Schlüter and H. Ney, "Model-based MCE bound to the true Bayes' error," *IEEE Signal Processing Letters*, vol. 8, no. 5, 2001.

[16] B. H. Juang and S. Katagiri, "Discriminative Learning for Minimum Error Classification," *IEEE Transactions on Signal Processing*, vol. 40, no. 12, pp. 3043–3053, 1992.

[17] G. Bouchard and B. Triggs, "The trade-off between generative and discriminative classifiers," in *Proc. COMPSTAT*, J. Antoch, Ed. Physica-Verlag, 2004.

[18] H. Chan and P. Woodland, "Improving broadcast news transcription by lightly supervised discriminative training," in *Proc. of ICASSP*, vol. 1, may 2004, pp. I – 737–40 vol.1.

[19] J. Zheng and A. Stolcke, "Improved discriminative training using phone lattices," in *Proc. of Interspeech*, 2005, pp. 2125–2128.

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