Lightly supervised acoustic model training for imprecisely and asynchronously transcribed speech

Conference Paper · October 2013

DOI: 10.1109/SpeD.2013.6682653

2 authors, including:

Péter Mihajlik
Budapest University of Technology and Eco…

42 PUBLICATIONS **198** CITATIONS

SEE PROFILE

Lightly Supervised Acoustic Model Training for Imprecisely and Asynchronously Transcribed Speech

Péter Mihajlik[1,2], András Balog[1]

[1] THINKTech Research Center
Vác, Hungary
{mihajlik,balog}@thinktech.hu

[2] Dept. of Telecom. & Media Informatics
Budapest Univ. of Technology and Economics
Budapest, Hungary

***Abstract*—In a variety of speech recognition tasks a large amount of approximate transcription is available for the audio material, but is not directly applicable for acoustic model training. Whereas roughly time-synchronous closed-captions or proper audiobook texts are already used in lightly supervised techniques, the utilization of more imperfect and at the same time completely unaligned transcriptions is not self-evident. In this paper we describe our experiments aiming at automated transcription of Hungarian parliamentary speeches. Essentially, a lightly supervised across-domain acoustic model adaptation/re-training is performed. A low-resource broadcast news model is used to bootstrap the process. Relying on automatic recognition of parliamentary training speech and on dynamic text alignment based data selection, a new, task-specific acoustic model is built. For the adaptation to the parliamentary domain, only edited official transcriptions and unaligned speech data are used, without any additional human annotation effort. The adapted acoustic model is applied on unseen target speech in real-time recognition. The word accuracy difference between the automatic and the human powered, official transcription is only 5% (as compared to the exact reference text).**

Keywords—large vocabulary continuous speech recognition; lightly supervised training; acoustic modeling; cross-domain adaptation

I. INTRODUCTION

Today's expectations from a Large Vocabulary Continuous Speech Recognition (LVCSR) system is human comparable accuracy, real-time operation and rapid development time. The fulfillment of these requirements can be extremely difficult, especially if low budget is envisaged. Fortunately, some recognition tasks abound in acoustic data and in their – approximate – transcriptions, such as Broadcast News with closed-captions [1, 2] or Audiobooks with faithful readings [3, 4]. Lightly supervised techniques are often applied for these tasks in order to utilize the tremendous human annotation effort – even if these transcriptions and recordings were originally not intended to serve speech technology. These lightly supervised techniques are typically applied on roughly time synchronous and/or relatively precise transcriptions.

Our task, the automated transcription of Hungarian parliamentary speeches, is similarly fortunate to the previous ones as the speeches and their official text contents are publicly available for the last decade [5]. The first challenge is the surprising inaccuracy of the official text (see Table II) as compared to the verbatim reference transcription. This phenomenon is not due to a sloppy typing but to the characteristics of Hungarian parliamentary speeches. On the one hand, Hungarian is a highly agglutinative, compounding language [6] for which the word error rate (WER) can be rather pessimistic [7]. On the other hand, spontaneous, quasi-spontaneous political debates are also common in the parliament. Hesitations, repetitions and agrammatical sentences occurring frequently in the speeches are corrected in the edited, official transcription. The other difficulty is that only the beginning and end of a speech is fixed approximately in time. No utterance- or lower level synchronization is available. The start-end times are inaccurate – several seconds differences to the actual times are observed resulting in speaker and text mismatch.

In the following, lightly supervised approaches and other related work are surveyed. Next, the Hungarian Parliament speech recognition task and data are described. Then we introduce a method for boosted lightly supervised acoustic modeling. Finally, the results obtained with lightly supervised acoustic model training are discussed and some conclusions are drawn.

II. RELATED WORK

One of the earliest work on lightly supervised acoustic model training was performed by Lamel and colleagues [1] aiming at automated *Broadcast News* (BN) transcription. In that study, beside written news text, closed-captions were integrated into the language model used to transcribe automatically the audio training corpora. Textual dynamic programming based data selection was used to select audio segments which match the captions, then a new acoustic model was built. For the bootstrap, a minimal amount of manually transcribed data was used, and iteratively, through light supervision, training data was broadened. The performance of the resulting acoustic model was comparable to a conventionally made one in terms of WER – without using substantial amount of additive manual annotation. The improvement due to training data selection was minor as compared to the unfiltered data set.

Reference [2] extended the previous work with discriminative training. The task was *BN recognition* too, and it was found that MPE (Minimum Phone Error) training outperforms the ML (Maximum Likelihood) method. Another conclusion was that the filtering of training data based either on matching closed-captions or confidence measurements had no real benefit.

In [3] lightly supervised technique was used to select precise readings of *audiobook* segments and align to the original text. For the LVCSR based selection a well-resourced acoustic model was used. The light supervision in the recognition was ensured by interpolating a language model trained on the book text with a background, English Gigaword based language model. With a minimal false acceptance rate the majority of the sentences could be selected. No acoustic model was trained on the selected data since the purpose of the work was TTS (Text-To-Speech) development. Reference [4] proposes a technique for almost automatic generation of acoustic models from unsynchronized audio and text material. The evaluation was performed on an Italian language *audiobook*. The technique assumes that the reader follows the book's text, thus no acoustic training data filtering is performed – the focus is on segmentation of large audio recordings.

Considering *parliamentary* speech recognition studies, conventional supervised, lightly supervised and unsupervised approaches can be found, as well. From *supervised* results we cite only the ones obtained on morphologically rich languages – that is merely Czech in our case. In [8] 100 hours of precisely (manually) transcribed parliamentary speech was used for supervised training of the acoustic models. On a test set, with a perplexity of 12.4 and OOV (Out Of Vocabulary) rate of 2.4%, a baseline WER of 17% could be achieved and was reduced further to 15% using speaker-cluster fusion methods. Reference [9] reports results on Czech language parliament speech with OOV rate of 1.1%, and WER of 20.8%.

Reference [10] introduces recognition results on Japanese parliamentary speech. It proposes a *lightly supervised* training scheme based on statistical language model transformation using MT (Machine Translation) technology, which fills the gap – that is 13% edit distance (or WER) in average – between faithful transcripts of spoken utterances and final texts for documentation. A consistently achieved character accuracy of nearly 90% is reported.

Finally, it must be added, that even an *unsupervised* approach to transcribe Polish language European Parliamentary speech automatically has been described in [11]. Similarly to the lightly supervised approaches, preliminary parliamentary speech transcriptions (0.5 million words) are indeed used in the language model – beside other Polish language EU documents and newspapers. This language model was then applied for both the automatic transcription of the acoustic model training set and for the actual recognition of the test set. The seed acoustic model was a Spanish language one mapped to Polish phonemes. Using confidence measure based training data filtering and various adaptation techniques iteratively, the performance reached a decent level (18% in terms of WER). In opposite to the previously cited approaches [1-4], it is not guaranteed here that the preliminary transcriptions belong to the audio training data – although, there is no evidence to the contrary, as well.

All in all, the effect of training data filtering is ambiguous in the related literature and across domain acoustic model training with light supervision is less investigated.

III. HUNGARIAN PARLIAMENT ASR TASK

The objective is to enable accurate real-time transcription of parliamentary speech so that the output can be used to aid official transcription, for closed-captioning for the disabled or index, search or analyze the content. In the current phase, the aim is to reach a reasonable real-time recognition performance (in terms of WER) using only readily available official transcriptions and sound recordings – as rapidly as possible.

The recordings in the sound archive of the Hungarian parliament are stored in variable bit-rate MP3 formats. We were able to access about 4000 hours of speech along with the edited, official transcriptions. All of the recordings were converted to 16kHz, 16 bit linear PCM (Pulse Code Modulation) format. Meta-data indicate the time boundaries of a speech segment in the audio stream. In the experiments, three distinct subsets were defined (Table I). The Train set was obtained by sampling randomly the original corpora. Segments with obviously erroneous meta-data were excluded, as well as the very long ones for computational reasons. To tune and evaluate the recognition system, the Tune and Eval subsets, respectively, were transcribed by professionals resulting in verbatim transcriptions.

The differences between official and verbatim transcriptions are illustrated in Table II ("sentence" here refers to variable long – up to several minutes – segments). The challenge in the acoustic modeling is not only due to this high level of textual disagreement but also to the large jitter of start and end times associated to the segments. Thus, a segment may actually contain speech from the preceding and/or the following speaker – and the actual speaker may be disrupted by the time boundaries. All in all, there is not a one-to-one mapping between the words of the audio and text segments.

For language modeling we used all the official transcriptions (Table III) excluding only the Tune and Eval texts. No verbatim transcription was used directly in the language or acoustic modeling.

TABLE I. ACOUSTIC CORPORA

	Tune	**Eval**	**Train**
Net Duration	4.7h	5.3h	500h
# Segments	131	164	28k
# Running words	32k	33k	3.4M
# Speakers	74	83	684

TABLE II. DISAGREEMENT (EDIT DISTANCE OR ERROR RATE) BETWEEN EDITED AND VERBATIM TRANSCRIPTIONS ON WORD AND SENTENCE LEVEL

	Word				**Sentence**
	Sub	**Del**	**Ins**	**Error**	**Error**
Tune	4.5%	3.6%	5.8%	13.9%	95.4%
Eval	4.5%	2.7%	6.1%	13.2%	92.1%

TABLE III. TRAINING TEXT CORPUS

	Train
# Running words	36.7M
# Unique words	571k

IV. LIGHTLY SUPERVISED TRAINING

Looking at Table II – especially at the sentence error rates – and at the segment boundary issue, no long explanation is required why simple forced alignment techniques are inapplicable. So, a natural choice is the application of Lightly Supervised (LS) techniques in order to produce coherent audio and text fragments for acoustic model training.

References [1, 2, 4, 11] propose several iterations in the cycle of automatic transcription of the training data, filtering, new acoustic model training/adaptation. However, these LS iterations can be computationally expensive and the improvements may be marginal only. In practice, already trained acoustic models are available for many languages. Not using such resources would be unfeasible in most situations. Our intention was to boost the LS training procedure with an out of domain bootstrap acoustic model and with an enhanced language modeling technique described in the next subsection.

A. *Dynamic language modeling through portion wise interpolation*

Typically a single static language model (LM) is used throughout the LS recognition of the acoustic training corpus [1, 2, 3, 11]. This language model is built on all the available transcriptions of the acoustic training data supplemented optionally with more in-domain text. The problem with this approach is that it does not take into account the actual part of the training data, even though its approximate text content is known. Therefore, – as the generalization of [3] approach – we defined portions of the training data, for which optimized language models can be used in the LS transcription, see (1). For each portion, T, a dedicated language model, LM_T, is built using the official transcriptions. This portion language model can be interpolated dynamically with a background language model, LM_{Bg}, trained on all training text (see Table III).

$$\hat{\lambda}_T = \underset{\lambda}{argmin}\, PP_T\left((1-\lambda)LM_T + \lambda\, LM_{Bg}\right) \qquad (1)$$

where PP is the word perplexity of the (interpolated) language model applied for LS transcription. For the portions of training data, about 5 hours of – preferably consecutive segments of – parliamentary speeches were chosen. Though it is possible to calculate and use custom interpolation weights for each portion of the training data, for simplicity, we applied the one optimized on the Tune set. Both with official and verbatim transcriptions an optimal value of around 0.2 was found for λ.

Word 3-gram models with modified, interpolated Kneser-Ney smoothing [12] were trained for both the background and the portion LM's using the SRILM toolkit [13]. The application of word language models for Hungarian though, may appear as inadequate. Nevertheless, this is a well resourced task, and according to our recent experiences, the application of subword lexical units in such circumstances comes with little or no benefit [14, 15].

Table IV shows various text statistics measured with different language models, where

- *normal* refers a LM trained on all independent training text,
- *static* stands for a LM trained on training + tune texts,
- and *dynamic* refers to the previously described LM, where the training+tune LM is linearly interpolated with the LM trained on tune portion's (official) text.

The perplexities and OOV rates were calculated both on the official and the verbatim transcriptions of the Tune and Eval sets.

It can be seen that dynamic language model interpolation radically reduces the test perplexities, promising better LS transcription accuracies, that may result in better acoustic models.

B. *Training data filtering*

Probably an oversimplified interpretation of LS training approaches is where the effectiveness is measured by the ratio of selected – presumably accurate – transcriptions. Yet this is a simple and comprehensible measure worth to look at it. The other – less investigated – measure is the "cleanness" of the selected data. Filtering may be based either on comparing the recognized text to the official one through the well-known dynamic programming method used to evaluate WER or on acoustic confidence measures. In this study we applied only the former one since we did have roughly accurate reference text – unlike in the unsupervised scenario which is not the subject of this work.

Table V and VI show the LS transcription results made with a Broadcast News (BN) acoustic model [14] with tight pruning (resulting in a Real-Time Factor, RTF=0.1) and the results of in-domain, 1st pass LS acoustic model (see Table VII) with loose pruning parameters (RTF~1), respectively. The net amount of data used actually in training is reduced further by filtering fragments shorter than 2 seconds (see Table VII for final selection values).

TABLE IV. STATISTICS WITH VARIOUS LANGUAGE MODELING APPROACHES ON VERBATIM AND OFFICIAL TEXTS

		Verbatim		Official	
Set	**LM type**	**PP**	**OOV**	**PP**	**OOV**
Eval	Normal	350	1.6%	262	0.7%
Tune	Normal	398	1.8%	306	0.8%
	LS Static	196	1.0%	122	0.0%
	LS Dynamic	78	1.0%	43	0.0%

It can be seen that dynamic language modeling achieved less improvements compared to the static one in terms of WER (and gross selection rate) as might be expected based on the perplexity tests. However, it allowed a similarly high level of training data filtering both in terms of quality and quantity with the less precise acoustic model than with the in-domain model using static LM.

TABLE V. WORD ERROR AND GROSS SELECTION RATES OF LIGHTLY SUPERVISED ALIGNMENT WITH OUT-OF-DOMAIN ACOUSTIC MODEL ON THE TUNE SET

LM type	Select	Sub	Del	Ins	Err
LS static	100%	15.0%	10.7%	3.3%	29.1%
	71%	1.9%	0.0%	1.4%	3.4%
LS dynamic	100%	9.1%	7.6%	2.8%	19.6%
	81%	2.2%	0.0%	1.5%	3.8%

TABLE VI. WORD ERROR AND GROSS SELECTION RATES OF LIGHTLY SUPERVISED ALIGNMENT WITH IN-DOMAIN ACOUSTIC MODEL ON THE TUNE SET

LM type	Select	Sub	Del	Ins	Err
LS static	100%	8.8%	3.2%	3.8%	15.8%
	84%	2.1%	0.0%	1.7%	3.7%
LS dynamic	100%	6.6%	3.1%	3.6%	13.3%
	87%	2.3%	0.0%	1.7%	4.0%

Note, that independently from the applied acoustic and language models, the selected training fragments contain similarly low amount of transcription errors.

C. *Acoustic model training on fragments*

Once the training data is filtered by keeping only those word sequences that match in the recognized text and the official transcriptions, and aligned to the corresponding audio fragments, the acoustic models can be trained. The only issue to be addressed is the handling of unknown phonetic contexts in fragment boundaries. Unlike [16], we do not trust the automatic or the official transcriptions outside the selected fragments, and therefore these contexts are treated as "junk" and are not used in the final acoustic model. Due to the inaccurate start and end time marking, the beginning and/or the end of the sound segment may contain words that aren't written in the official text. To allow these words in the forced alignment, a compound "non-sil" acoustic model was synthesized of all the monophones, and was used to match the extra words during the alignment.

V. RECOGNITION RESULTS

A. *Recognition setup*

Throughout the experiments the setup was as follows. Standard MFCC (Mel Frequency Cepstral Coefficient) features with blind channel equalization [17] were extracted, including first and second derivatives and energy, resulting in a total of 39 dimensional vectors. Across-word, 3-state, left-to-right structure shared-state triphone HMM (Hidden Markov-Model) acoustic models were trained with GMM(Gaussian Mixture Model) density functions. At the end of the ML training process, about 5000 tied states with 10 Gaussians per state were produced for each acoustic model, using the HTK toolkit [18].

Word lexical modeling was applied with *full* vocabulary of 571k words. Phonemic pronunciations were generated automatically based on grapheme-to-phoneme rules and with the application of a small exception dictionary for abbreviations, acronyms, etc.

In each recognition turn on the Eval set, the same 3-gram language model was used. The LM was trained on the full Train set – excluding test texts – with the modified Kneser-Ney smoothing.

The above speech recognition knowledge sources were integrated and optimized in the WFST (Weighted Finite State Transducer) framework [19]. In all recognition experiments the VOXerver decoder [6, 14] was used on a 3.4GHz CPU enabling faster than real-time operation. The pruning settings were adjusted to nearly saturate the accuracy-RTF curve, typically resulting in a RTF of 0.2 (5x faster than real time). The various acoustic models were evaluated on the unseen Eval set – overlap between Train and Eval set speakers is presumable, but not checked.

B. *Experimental results*

First, classical lightly supervised experiments were conducted. For initialization, an acoustic model was trained in the classical, supervised manner on the 500 hours of parliamentary speech data with the official rough transcriptions. Then, using the resulted acoustic model with a static language model, a LVCSR pass on the training data was performed. After filtering out the nonmatching segments a new – 1st LS pass – acoustic model was built and evaluated on the Eval set. The results of several LS iterations are shown in Table VII. It can be seen, that the first LS training achieves great improvement but the further iterations are less successful.

TABLE VII. CONVENTIONAL LS RESULTS WITH FILTERING

Acoustic model	WER	In-domain data
Supervised with rough transcription	27.0%	500h
1st LS pass (static LM)	19.5%	329h
2nd LS pass (static LM)	19.1%	391h
3rd LS pass (static LM)	19.3%	359h

TABLE VIII. ADAPTATION RESULTS

Acoustic model	WER	In-domain data
BN (baseline, no adaptation)	26.3%	-
MLLR supervised with official text	23.5%	500h
MLLR unsupervised	21.3%	500h
MLLR unsupervised + Filtering	21.3%	344h
MAP unsupervised + Filtering	20.1%	344h
MLLR+MAP unsupervised + Filtering	19.3%	344h

TABLE IX. CROSS-DOMAIN LS RESULTS

Acoustic model	WER	In-domain data
LS + Static LM interpolation	19.3%	500h
LS + Static LM interpolation + Filtering	18.6%	344h
LS + Dynamic LM interpolation	19.1%	500h
LS + Dynamic LM interpolation + Filtering	18.3%	373h

The next approach was to apply an out of domain acoustic model (BN) – which was trained on 33 hours of broadcast news speech data [14, 15] – and perform conventional acoustic model adaptation. The first adaptation attempt was to apply the official transcriptions along with MLLR (Maximum Likelihood Linear Regression). The second, unsupervised adaptation experiment gave significantly better results. The filtering of the adaptation data and the application of MLLR+MAP adaptation, as well, achieved the best results (see Table VIII).

Finally, the effect of static and dynamic language model interpolation based filtering was investigated in the LS (Lightly Supervised) framework when initializing with the BN acoustic model. As Table IX demonstrates, the effect of filtering is clearly beneficial, however, the dynamic LM interpolation approach achieves less improvement as one would expect based on preliminary results of Table IV, V, VI. Still, the overall best results is obtained with only one pass LS training with the proposed, out of domain acoustic model initialization and dynamic language model interpolation for the LVCSR-based training data filtering.

VI. Conclusions

In this study an assortment of lightly supervised acoustic modeling techniques were reviewed. Based on the experiments on a Hungarian Parliamentary speech recognition task, some – hopefully useful – conclusions could be drawn. First, as Table IX shows, using the proposed dynamic language modeling approach and an out-of-domain bootstrap acoustic model, the lightly supervised training can be efficiently boosted. Only one pass training was enough to achieve the best results on a given training set. Second, in our setup, unlike in [2], filtering of training data reduced consistently and significantly (more than 5%, relative) the error rates. Another finding is that when filtering, neither the quality of the applied acoustic model nor the perplexity of the language correlates with the resulted training transcription accuracy – which was high enough in each case. The model qualities apparently have a more direct impact on the ratio of selected data – that seems less crucial in our task.

In sum, the final recognition results – achieved without the need of any in-domain verbatim transcription – on the Hungarian parliamentary task seem to be comparable not only to the supervised results of other morphologically rich languages but nearly – in a numerical sense – with the official human transcriptions.

As for future work, we plan to apply various speaker adaptive and discriminative training techniques.

Acknowledgment

Our research was partially funded by the following national projects/grants by NFU: KMR 12-1-2012-0207 (DIANA), GOP-1.1.1-11-2012-0377 (WEBRA TimeSave).

References

[1] Lamel, L., Gauvain, J.-L., and Adda, G.. "Lightly supervised acoustic model training." In Proc. ASR2000-Automatic Speech Recognition: Challenges for the new Millenium ISCA Tutorial and Research Workshop (ITRW), Paris, France, pp. 150-154, 2000.

[2] Chan, H. Y., & Woodland, P., "Improving broadcast news transcription by lightly supervised discriminative training". In Proceedings IEEE International Conference on Acoustics, Speech, and Signal Processing.(ICASSP'04). Vol. 1, pp. I-737, 2004.

[3] Braunschweiler, N., Gales, M. J., & Buchholz, S. "Lightly supervised recognition for automatic alignment of large coherent speech recordings". In Proceedings of the 11th Annual Conference of the International Speech Communication Association, pp. 2222-2225, 2010.

[4] Alessandrini, M., Biagetti, G., Curzi, A., and Turchetti, C., "Semi-automatic acoustic model generation from large unsynchronized audio and text chunks." In Twelfth Annual Conference of the International Speech Communication Association., Florence, Italy, pp. 1681-1684, 2011.

[5] http://www.parlament.hu (Official website of the Hungarian Parliament)

[6] Mihajlik, P., Tüske, Z., Tarján, B., Németh, B., Fegyó, T., "Improved Recognition of Spontaneous Hungarian Speech – Morphological and Acoustic Modeling Techniques for a Less Resourced Task", *IEEE Trans Audio Speech & Lang. Proc.*, vol.18, no.6, pp.1588-1600, Aug. 2010.

[7] Creutz, M., Hirsimäki, T., Kurimo, M., Puurula, A., Pylkkönen, J., Siivola, V., Varjokallio, M., Arisoy, E., Saraçlar M., and Stolcke, A. "Morph-based speech recognition and modeling of out-of-vocabulary words across languages". *ACM Transactions on Speech and Language Processing*, vol. 5, no 1, 2007:3.

[8] Vaněk, J., Psutka, J.V., Zelinka, J., Trmal, J., "Training of Speaker-Clustered Discriminative Acoustic Models for Use in Real-Time Recognizers", Speech Processing, vol. 2010, pp. 152-158, Institute of Photonics and Electronics AS CR, Prague, 2010.

[9] Nouza, Jan, Jindrich Zdansky, Petr Cerva, and Jan Silovsky. "Challenges in speech processing of Slavic languages (case studies in speech recognition of Czech and Slovak)." *Development of Multimodal Interfaces: Active Listening and Synchrony*, pp. 225-241, 2010.

[10] Kawahara, T., "Transcription System Using Automatic Speech Recognition for the Japanese Parliament (Diet)", In Proceedings of the Twenty-Fourth Innovative Appications of Artificial Intelligence Conference, pp. 2224-2228, 2012.

[11] Lööf, J., Gollan, C., and Ney, H., "Cross-language Bootstrapping for Unsupervised Acoustic Model Training: Rapid Development of a Polish Speech Recognition System", In Proc. Int. Conf. on Spoken Language Processing, Brighton, UK, pp. 1617–1620, 2009.

[12] Chen, S. F., and Goodman, J.T., "An Empirical Study of Smoothing Techniques for Language Modeling" Technical Report TR-10-98, Computer Science Group, Harvard University, 1998.

[13] Stolcke, A., "SRILM – an extensible language modeling toolkit", in Proc. Intl. Conf. on Spoken Language Processing, pp. 901–904, Denver, 2002.

[14] Tarján, B., Mihajlik, P., Balog, A., Fegyó, T., "Evaluation of Lexical Models for Hungarian Broadcast Speech Transcription and Spoken Term Detection." In: 2011 2nd International Conference on Cognitive Infocommunications, CogInfoCom 2011. Budapest, Hungary, 2011.

[15] Tarjan, B., T. Mozsolics, A. Balog, D. Halmos, T. Fegyo, and P. Mihajlik. "Broadcast news transcription in Central-East European languages." In Proc. 3rd International IEEE Conference on Cognitive Infocommunications (CogInfoCom), IEEE, pp. 59-64, 2012.

[16] Gollan, C., Hahn, S., Schlüter, R., & Ney, H. (). "An improved method for unsupervised training of LVCSR systems". Interspeech, Antwerp, Belgium, pp. 2101-2104, 2007.

[17] Mauuary, L., "Blind Equalization in the Cepstral Domain for robust Telephone based Speech Recognition", Proc. EUSPICO'98, Vol.1, pp. 359-363, 1998.

[18] Young, S. et al., The HTK book. (for HTK version 3.4.), 2006

[19] Mohri, M., Pereira, F. and Riley, M., "Weighted Finite-State Transducers in Speech Recognition", *Computer Speech and Language*, 16(1), pp. 69-88, 2002.