LANGUAGE MODEL ADAPTATION FOR ASR OF SPOKEN TRANSLATIONS USING PHRASE-BASED TRANSLATION MODELS AND NAMED ENTITY MODELS

Joris Pelemans[1]*, Tom Vanallemeersch*[2]*, Kris Demuynck*[3]*,*
Lyan Verwimp[1]*, Hugo Van hamme*[1]*, Patrick Wambacq*[1]

[1]ESAT, KU Leuven, Belgium
[2]Centre for Computational Linguistics, KU Leuven, Belgium
[3]ELIS, Ghent University, Belgium

{joris.pelemans,lyan.verwimp,hugo.vanhamme,patrick.wambacq}@esat.kuleuven.be
tom.vanallemeersch@ccl.kuleuven.be
kris.demuynck@elis.ugent.be

ABSTRACT

Language model adaptation based on Machine Translation (MT) is a recently proposed approach to improve the Automatic Speech Recognition (ASR) of spoken translations that does not suffer from a common problem in approaches based on rescoring i.e. errors made during recognition cannot be recovered by the MT system.

In previous work we presented an efficient implementation for MT-based language model adaptation using a word-based translation model. By omitting renormalization and employing weighted updates, the implementation exhibited virtually no adaptation overhead, enabling its use in a real-time setting.

In this paper we investigate whether we can improve recognition accuracy without sacrificing the achieved efficiency. More precisely, we investigate the effect of both state-of-the-art phrase-based translation models and named entity probability estimation. We report relative WER reductions of 6.2% over a word-based LM adaptation technique and 25.3% over an unadapted 3-gram baseline on an English-to-Dutch dataset.

Index Terms: speech recognition, spoken translations, language model adaptation, phrase-based machine translation, named entities

1. INTRODUCTION

Although computer-aided translation (CAT) is traditionally performed with keyboard and mouse, a recent study [1] has shown that in the context of machine translation (MT) the use of automatic speech recognition (ASR) as an input method may constitute a significant speed-up, even with a non-perfect speech transcription that needs additional correction. Furthermore, it has been established that by using the translation model (TM) and more specifically the translation probabilities of the words and/or word groups (phrases) of the source language text, the speech recognition of spoken translations can be improved [2, 3, 4, 5, 6].

There are several different scenarios that can be applied to combine models for decoding the optimal transcription of a spoken translation where each implies different assumptions about system implementation and constraints on computational complexity. All of them are based on a Bayesian extension of the ASR maximum likelihood formula, first proposed by [2]: given a source language text $F = f_1...f_J$ and an acoustic signal $X = x_1...x_T$, which is the spoken version of a target language text $E = e_1...e_I$, the optimal transcription $\hat{E}$ is decoded as follows:

$$\begin{aligned}\hat{E} &= \underset{E}{\operatorname{argmax}}\, P(E|X,F) = \underset{E}{\operatorname{argmax}}\, P(X,F|E)P(E) \\ &= \underset{E}{\operatorname{argmax}}\, P(X|E)P(F|E)P(E) \end{aligned} \quad (1)$$

where the conditional independence assumption of X and F given E is considered to be reasonable and allows a decomposition into three knowledge sources: the acoustic model $P(X|E)$, the translation model $P(F|E)$ and the language model $P(E)$.

One scenario in which this extension can be used is to rescore hypotheses generated by the ASR system, using the TM probabilities, as part of a multi-pass approach. This can be done either by reranking ASR n-best lists [7, 8, 3] or by rescoring word lattices [4, 5]. One of the main issues with multi-pass approaches, is that the recognizer does not have access to MT information during the first pass. The recognizer might have already pruned out several interesting hypotheses that no rescoring can ever recover. Another issue is that the output of the recognizer has to be stored and that the second pass can only start when the first pass has finished, which takes up valuable space and time. In human-computer interaction, response times and storage should be minimized to reduce the overhead associated with rescoring.

Only very recently did the multi-pass scenario make room for a new scenario which we believe will be the new paradigm in MT-ASR integration. [6] shows that, for each source language sentence, the language model of the ASR system can be updated using information derived from the text of that individual sentence. This approach is claimed to yield a decrease in computational complexity by efficiently pruning the LM as well as a significant reduction in word error rate (WER).

In previous work [9] we showed that the efficiency of updating a LM according to the techniques described in [6] can still be greatly improved and we proposed a method that allows for LM adaptation without renormalization and with limited extra storage and fewer and on-the-fly updates.

In this paper we investigate whether we can improve recognition accuracy without sacrificing the achieved efficiency. More precisely, we investigate the effect of both state-of-the-art phrase-based translation models and named entity probability estimation.

This research is funded by the Flemish government agency IWT (project 130041, SCATE).

In the remainder of the paper we discuss related work (Section 2), phrase-based translation models (Section 3) and named entity probability estimation (Section 4), and validate our work experimentally (Section 5). We end with conclusions and future work in Section 6.

2. RELATED WORK

2.1. Word-based Language Model Adaptation

In [9] we proposed an efficient implementation of a technique proposed by [6]. Instead of applying a multi-pass approach, the authors propose a direct integration of the translation model probabilities into the ASR language model. They do this by approximating $P(F|E)$ in Eq. 1 at the sentence level, only taking into account lexical translation probabilities. That is, each word e_i in the target language sentence E corresponds to exactly one word f_j in the source language sentence F:

$$P(F|E) \approx \prod_{i=1}^{I} \max_{j:f_j \in F} P(f_j|e_i) \tag{2}$$

The product of $P(F|E)$ and $P(E)$ can then be considered as the new language model $P'(E)$ with which speech is decoded.

If we consider an n-gram to be a tuple (h, e) consisting of a history h and a target word e, then adapting the n-gram probability $P(e|h)$ to $P'(e|h)$ boils down to:

$$P'(e|h) = \frac{P(e|h) \max_{j:f_j \in F} P(f_j|e)}{norm(h)} \tag{3}$$

The normalization factor $norm(h)$ is obtained as follows:

$$norm(h) = \frac{\sum_{e':(h,e') \in \mathcal{T}} P(e'|h) \max_{j:f_j \in F} P(f_j|e')}{\sum_{e':(h,e') \in \mathcal{T}} P(e'|h)} \tag{4}$$

where $\mathcal{T}$ corresponds to the training data used to train the original language model probabilities $P(e|h)$.

Once all relevant probabilities have been updated, the back-off weights $bow(h)$ are also renormalized to obtain a true probability distribution $P'(e|h)$:

$$bow(h) = \frac{1 - \sum_{e':(h,e') \in \mathcal{T}} P'(e'|h)}{1 - \sum_{e'':(h,e'') \notin \mathcal{T}} P'(e''|h)} \tag{5}$$

2.2. Efficient Implementation

In [9] we argued that the approach described in Section 2.1 is suboptimal wrt computational complexity and disk storage. One issue is that all the n-grams need to be updated, even the irrelevant n-grams i.e. the ones for which $P(f|e)$ is absent from the TM. This is because not updating these n-grams corresponds to multiplying the original probability $P(e|h)$ by 1 which is always larger than or equal to $\max_{j:f_j \in F} P(f_j|e)$, leading to relatively larger probabilities for irrelevant n-grams. Therefore it is necessary to decrease the probabilities of the irrelevant n-grams e.g. by multiplying them with a small value $\epsilon \ll 1$, but then each n-gram in the model needs updating, which has a negative impact on efficiency.

More importantly however, the approach requires the computation of the normalization factor and back-off weight for each history h which is very time consuming. The end result is a new LM for each sentence, which requires a lot of storage and for which the speech recognizer needs to switch language models every time.

Our work was based on the observation that LM normalization is not mandatory. In the context of ASR decoding, a score is not attributed to a single word sequence, but rather to many competing word sequence hypotheses. Therefore, there is no strict constraint that the score that a LM attributes to each hypothesis should obey a true probability distribution, as long as a more likely hypothesis receives a higher score. For a more elaborate discussion on this, we refer the reader to [9].

To reduce the number of n-gram updates we also introduced the notion of weighted updates. Rather than multiplying n-gram probabilities directly with translation model probabilities $P(f|e) \leq 1$, we instead multiply the relevant n-gram probabilities by weights $g(f, e) \geq 1$. This means that the irrelevant ones can remain untouched, thus significantly reducing the number of updates. We proposed an alternative update rule that essentially inflates rather than deflates the n-gram probabilities:

$$score_{LM}(e|h) = P(e|h) \max_{j:f_j \in F} g(f_j, e) \tag{6}$$

where g is a weighting function that maps TM probabilities $P(f|e)$ onto values larger than 1. To have some control over the shape and maximum of this function, yet minimize the introduction of new parameters, we proposed the following exponential function:

$$g(f, e) = 1 + \alpha\beta^{1-P(f|e)} \tag{7}$$

where $\alpha \in \mathbb{R}_0^+$ controls the maximum value of the update weight and $\beta \in]0, 1[$ determines the relative weight that is given to $P(f|e)$: a smaller value of β will give a relatively higher weight to high probabilities than to low probabilities. It is important that these parameters are carefully optimized, because if we inflate the LM probabilities too much we run the risk of overshadowing the acoustic score or the word insertion cost and ending up with acoustically implausible hypotheses or hypotheses consisting only of short words.

Finally we reduced disk storage and memory usage by storing only the inflation weights, rather than the updated LM in its entirety.

3. PHRASE-BASED LANGUAGE MODEL ADAPTATION

3.1. Phrase-based Machine Translation

Although the implementation described in Section 2.2 drastically improves the efficiency of MT-based LM adaptation, Eq. 2 assumes translation consists solely of one-to-one alignments i.e. each word f_j in the source language text can only correspond to one word e_i in the target language text. This is a strong assumption that does not hold in reality: every language has its own way of verbalizing concepts with some using a single word and others using multiple words for the same concept. In machine translation this issue is addressed by so-called phrase-based translation models [10].

Phrase-based translation models are models that cover correspondences between sequences of words, called phrases, which we denote by $\bar{f}$ and $\bar{e}$ for source and target language, respectively. This approach has several advantages. It allows for translating a longer sequence to a shorter sequence or vice versa. For instance, English *grey horse* translates to Dutch *schimmel*, and the English compound *screen resolution* to *schermresolutie*. It allows for capturing local context. For instance, *large horse* can be translated word by word to Dutch (*groot paard*), but the combination of *grey* and *horse* should be translated as a whole, as it (most likely) indicates a specific type of horse for which there is a specific word in Dutch. It should be stressed here that a phrase is not a linguistic notion in a phrase-based

MT model: the sequence *looks at the* (*kijkt naar de*) may be a phrase, although it is not a linguistic constituent. This example sequence also shows that phrases capture the context of words which have many possible translations, like prepositions. In the phrase *looks at the*, the word *at* should be translated into Dutch as *naar*; compare this to *laughs at the* (*lacht om de*) where *at* is translated as *om*.

Correspondences between phrases are constructed from word-aligned sentence pairs, and stored in a phrase table, i.e. a list of phrase pairs with associated scores. This is a three-step process, the details of which are discussed in the below paragraphs. In the first step, the two word alignments (one for each language direction) of a sentence pair are turned into a single, symmetrized one. In the second step, phrase pairs are extracted from the symmetrized word alignment of a sentence pair. In the third step, phrase pair scores are calculated based on the full set of phrase pairs extracted from all sentence pairs.

The first step in building a phrase table consists of combining the two word alignments of each sentence pair. The first alignment links source words to target words (a source word may be linked multiple times, i.e. have links with more than one target word). The second alignment links target words to source words (multiple linking is also possible here). The two word alignments are symmetrized, which can take place in several ways. For instance, if symmetrization only keeps links which occur in both alignments, we obtain an intersective word alignment. Another type of symmetrization uses a heuristic to add additional alignment points to the intersective alignment. From the symmetrized alignments of all sentence pairs, lexical probabilities $w(f|e)$ and $w(e|f)$ are estimated by relative frequency:

$$w(f|e) = \frac{count(f,e)}{\sum_{f'} count(f',e)} \tag{8}$$

where $count(f,e)$ stands for the number of times f aligns with e in the training data. These lexical probabilities express the likelihood that one word is translated by another one[1]. Their use is further detailed below.

The second step in the phrase table construction procedure consists in extracting consistently aligned phrase pairs from the symmetrized alignment of a sentence pair. These are phrase pairs in which each source word is either aligned to a word in the target phrase, or not aligned at all, and in which each target word is either aligned to a word in the source phrase, or not aligned at all.

The third and final step in creating the phrase table consists of calculating four scores for each phrase pair $(\bar{f},\bar{e})$. The first score is the phrase translation probability $\phi(\bar{f}|\bar{e})$, which is the phrase equivalent of the lexical probability $w(f,e)$ and is also estimated by relative frequency. The second score is the lexical weight, which validates the quality of a phrase pair by checking how well its words translate to each other. This serves to compensate for an overestimation of the reliability of rare phrase pairs. Given a phrase pair $(\bar{f},\bar{e})$ and a word alignment a between the target language phrase positions $i = 1,\ldots,m$ and source language phrase positions $j = 1,\ldots,n$ the lexical weight P_w is computed by:

$$P_w(\bar{f}|\bar{e},a) = \prod_{j=1}^{n} \frac{1}{|\{i:(i,j)\in a\}|} \sum_{\forall(i,j)\in a} w(f_j|e_i) \tag{9}$$

The third and the fourth score are inverses of the first and second one, i.e. they apply to the translation in the other direction.

[1]Before calculating these probabilities, a NULL token is added to the target sentence, which is aligned to each source word that was not aligned to a target word. This allows for calculating the probability that a source word is not translated. The same is done in the other direction.

The four phrase pair scores described above are combined by the MT system using a log-linear model, which also involves other scores, such as a language model score, a reordering score etc. We will not discuss these other scores further, as they are not relevant to the combination of phrases and ASR described in the present paper. The weights for the scores are determined during a tuning process where the MT output for a set of held-out source sentences is compared to their reference translations. Finally, when the MT system translates a source sentence F, it selects the target sentence E with the highest score according to the log-linear model.

3.2. Language Model Adaptation

By using phrase-based translation models, we can extend Eq. 2 as follows:

$$P(F|E) \approx \max_{a,\bar{f}\in F,\bar{e}\in E} \prod_{i=1}^{l} P(\bar{f}|\bar{e}_i) \tag{10}$$

where l denotes the number of phrases in E, given the alignment a that maximizes $P(F|E)$.

$P(\bar{f}|\bar{e})$ can then be estimated by combining the four phrase pair scores described in Section 3.1. Because the update weights in Eq. 7 are based on probabilities and the normalization of a log-linear model is expensive, we opted to combine the scores via linear interpolation:

$$\begin{aligned} P(\bar{f}|\bar{e}) = {} & \lambda_1\phi(\bar{f}|\bar{e}) + \lambda_2 P_w(\bar{f}|\bar{e}) \\ & + \lambda_3\phi(\bar{e}|\bar{f}) + \lambda_4 P_w(\bar{e}|\bar{f}) \end{aligned} \tag{11}$$

where the λ_i's are estimated empirically.

The weighting scheme allows us to reduce the number of updates, thus maintaining efficiency, provided that the amount of phrases to be evaluated is not too large. We will come back to this in Section 5.

4. NAMED ENTITIES

Many named entities are written in a similar or even identical form in multiple languages, especially if the languages belong to the same language family. This means that, even if the translation model does not contain a translation for a given named entity, there is a relatively high probability that the translation can just be copied from the source language to the target language. Choosing a value for the named entity translation probability $P(NE_f|NE_e) \approx 1$ then allows updates to n-grams containing this named entity, provided that it exists in the ASR language model and lexicon.

If however the named entity does not occur in the ASR language model and lexicon, there are no n-gram probabilities to update and we have to resort to language model estimation techniques. One simple estimation technique that we have found to work well is to map new named entities in the language model to a special token <UNK> for out-of-vocabulary (OOV) words. The n-gram probabilities for this special token are estimated on words that occur in the training data, but are excluded from the vocabulary. The named entities can then use the OOV probabilities, optionally weighted by a factor h_{NE}. If the pronunciation of the named entity is not too different in the source language and the target language, and the g2p converter is able to approximate the pronunciation in the source language, this very rough LM estimate often steers the recognizer into the right direction.

Note that the techniques described in this section are not novel by themselves, but as far as we know, their effect has not been investigated on MT-based LM adaptation using update weights and a phrase-based translation model.

5. EXPERIMENTS

5.1. Task and Setup

The ASR experiments were performed on a test corpus of audio fragments from the Flemish part of the Corpus Spoken Dutch (CGN) [11], component o. The chosen fragments correspond to 167 Dutch utterances that are read translations from English books.

Using a vocabulary of 100k words, an initial 3-gram LM with modified Kneser-Ney smoothing [12, 13] was trained by running the SRILM toolkit [14] on a collection of normalized newspaper texts from the Flemish digital press database Mediargus which contains 1104M word instances (tokens) and 5M unique words (types). The vocabulary was converted into a phonemic lexicon using an updated version of the g2p described in [15] and integrated into the recognizer described in [16], which was built with our in-house speech recognition system SPRAAK [17].

The TM was created by applying the GIZA++ toolkit [18] to a set of 1M English-Dutch parallel sentence pairs extracted from the Europarl corpus [19], which contains the written version of speeches of members of the European Parliament. GIZA++ adopts an EM approach to learning lexical probabilities $P(f|e)$ and $P(e|f)$ from a parallel corpus. The approach initializes the lexical probabilities using a uniform translation distribution for words. Based on this initialization, the probabilities of possible word alignments of sentence pairs are calculated. The new probabilities then allow to recalculate the lexical probabilities across the set of sentence pairs.[2] This process continues until convergence.

The phrase table was created using the process described in Section 3.1 (the first step added additional alignment points after calculating intersection) and filtered to contain only one-to-one and many-to-one alignments. Though many-to-many alignments could in principle yield further improvements, this comes at the cost of efficiency as the many-to-many alignments take up more than 90% of the phrase table. The final phrase table contains 2,939,355 phrase pairs which is ca. 1.6 times the size of the word-based translation model. As the computation of the update weights took ca. 0.2s per sentence for the word-based model, the overhead introduced by using a phrase-based model is negligible.

To build the updated LMs, we excluded all updates for source language words shorter than 4 letters, as we found that these tend to be unreliable. Words that have at least 2 letters and start with a capital letter are considered named entities and are never excluded. Named entities that are not in the TM, but are in the ASR lexicon, were given an optimized translation score of 1.2. Named entities that are not in the ASR lexicon were added to the lexicon using the g2p; their LM probabilities correspond to the OOV probabilities, weighted by $h_{NE} = 1.65$ which was optimized empirically.

Best results were achieved with phrase translation probabilities $\phi(\bar{f}|\bar{e})$ only i.e. $\lambda_1 = 1$, $\lambda_2 = 0$, $\lambda_3 = 0$ and $\lambda_4 = 0$. The inflation weights based on these probabilities were generated with optimized values of $\alpha = 16$ and $\beta = 0.0005$.

[2]For the sake of clarity, we would like to point out that these probabilities are different from the ones that are calculated from word alignment in the first step of phrase table construction.

	WER	Reduction
Initial LM (no TM updates)	25.96	-
Word-based adaptation	20.68	20.3%
Phrase-based adaptation	20.43	21.3%
+ named entities (IV)	20.19	22.2%
+ named entities (IV+OOV)	19.39	25.3%

Table 1. WERs (in %) and relative reductions using the investigated types of adaptation, compared to an initial 3-gram model that does not use TM probabilities. A distinction is made between in-vocabulary (IV) and out-of-vocabulary (OOV) named entities. The models are evaluated on 167 utterances from the Dutch CGN corpus (component o), corresponding to translations from English.

5.2. Results

Table 1 shows the word error rates (WERs) and relative reductions using the investigated types of adaptation, compared to an initial 3-gram model that does not use TM probabilities. It can be seen that the use of phrase-based rather than word-based translation models reduces the WER by 1% relatively. We suspect that this moderate reduction is in part due to filtering out many-to-many alignments in favor of adaptation efficiency. On the other hand it is probably also related to the nature of the test data: written stories often use complex structure and vocabulary which causes the translations to deviate substantially.

Simulating translation probabilities for English named entities that do not occur in the translation model, but do occur in the Dutch ASR lexicon and LM, also has a small, but significant effect on the recognizer. This confirms our intuition that the translation can just be copied from the source language to the target language, which is helped by the fact that in our case the source and target language are both Germanic languages.

Finally, we observe that even English named entities that are unknown to both the MT and ASR system, can be modeled well by using a Dutch g2p for the pronunciation and weighted OOV probabilities for the language model. Although it is to be expected that adding relevant words to the recognizer yields improvement, it is interesting that this constitutes the largest improvement of the investigated techniques, giving a total relative WER reduction of 6.2% compared to the word-based model and 25.3% over the unadapted 3-gram baseline.

6. CONCLUSIONS AND FUTURE WORK

We presented extensions on our efficient MT-based language model adaptation technique for automatic speech recognition of spoken translations. We investigated the effect of phrase-based translation model and named entity probability estimation and found that together they achieve a relative WER reduction of 6.2% over a word-based LM adaptation technique and 25.3% over an unadapted 3-gram baseline. Moreover, the extensions come with the same efficiency benefits as the word-based model which allow their use in a real-time CAT environment. To our knowledge this is the first MT-based language model adaptation technique using a phrase-based translation model.

In the future we plan to investigate on-the-fly calculation of phrase translation probabilities given source phrases of sentences [20]. This may allow us to retrieve and select many-to-many phrase pairs in a reasonable amount of time.

7. REFERENCES

[1] B. Dragsted, I.M. Mees, and I. Gorm Hansen, "Speaking your Translation : Students' First Encounter with Speech Recognition Technology," *Translation & Interpreting*, vol. 3, no. 1, pp. 10–43, 2011.

[2] P. Brown, S. Chen, S. Della Pietra, V. Della Pietra, S. Kehler, and R. Mercer, "Automatic Speech Recognition in Machine Aided Translation," in *Computer Speech and Language*, 1994, vol. 8, pp. 177–187.

[3] Matthias Paulik, Christian Fügen, Sebastian Stüker, Tanja Schultz, Thomas Schaaf, and Alex Waibel, "Document Driven Machine Translation Enhanced ASR," in *Proc. Interspeech*, 2005, pp. 2261–2264.

[4] Shahram Khadivi and Hermann Ney, "Integration of Automatic Speech Recognition and Machine Translation in Computer-assisted translation," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 16, no. 8, pp. 1551–1564, 2008.

[5] Aarthi M. Reddy and Richard C. Rose, "Integration of Statistical Models for Dictation of Document Translations in a Machine-Aided Human Translation Task," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 18, no. 8, pp. 2015–2027, 2010.

[6] Luis Rodríguez, Aarthi M. Reddy, and Richard C. Rose, "Efficient Integration of Translation and Speech Models in Dictation Based Machine Aided Human Translation," in *Proc. ICASSP*, 2012, pp. 4949–4952.

[7] J. Brousseau, G. Foster, P. Isabelle, R. Kuhn, Y. Normandin, and P Plamondon, "French Speech Recognition in an Automatic Dictation System for Translators: the TransTalk Project," in *Proc. Eurospeech*, 1995, pp. 193–196.

[8] Shahram Khadivi, Andr'as Zolnay, and Hermann Ney, "Automatic Text Dictation in Computer-Assisted Translation," in *Proc. Interspeech*, 2005, pp. 2265–2268.

[9] Joris Pelemans, Tom Vanallemeersch, Kris Demuynck, Hugo Van hamme, and Patrick Wambacq, "Efficient Language Model Adaptation for Automatic Speech Recognition of Spoken Translations," in *Proc. Interspeech*, 2015, pp. 2262–2266.

[10] Philipp Koehn, Franz Josef Och, and Daniel Marcu, "Statistical Phrase-Based Translation," in *Proc. HLT-NAACL*, 2003, pp. 48–54.

[11] Nelleke Oostdijk, "The Spoken Dutch Corpus. Overview and first evaluation," in *Proc. LREC*, 2000.

[12] Reinhard Kneser and Hermann Ney, "Improved Backing-off for M-gram Language Modeling," in *Proc. ICASSP*, 1995, vol. I, pp. 181–184.

[13] Stanley F. Chen and Joshua Goodman, "An Empirical Study of Smoothing Techniques for Language Modeling," 1998.

[14] Andreas Stolcke, "SRILM - An Extensible Language Modeling Toolkit," in *Proc. ICSLP*, 2002, pp. 257–286.

[15] Kris Demuynck, Tom Laureys, and Steven Gillis, "Automatic Generation of Phonetic Transcriptions for Large Speech Corpora," in *Proc. ICSLP*, 2002, pp. 333–336.

[16] Kris Demuynck, Antti Puurula, Dirk Van Compernolle, and Patrick Wambacq, "The ESAT 2008 System for N-Best Dutch Speech Recognition Benchmark," in *Proc. ASRU*, 2009, pp. 339–343.

[17] Kris Demuynck, *Extracting, Modelling and Combining Information in Speech Recognition*, Ph.D. thesis, K.U.Leuven ESAT, 2001.

[18] Franz Josef Och and Hermann Ney, "A Systematic Comparison of Various Statistical Alignment Models," *Computational Linguistics*, vol. 29, no. 1, pp. 19–51, 2003.

[19] Philipp Koehn, "Europarl: A Parallel Corpus for Statistical Machine Translation," in *Proc. Machine Translation Summit*, 2005, pp. 79–86.

[20] Chris Callison-Burch, Colin Bannard, and Josh Schroeder, "Scaling Phrase-based Statistical Machine Translation to Larger Corpora and Longer Phrases," in *Proc. ACL*, 2005, pp. 255–262.