

# Deriving Phrase-based Language Models<sup>1</sup>

**Peter A. Heeman**

Center for Spoken Language Understanding  
Oregon Graduate Institute  
PO Box 91000 Portland Oregon 97291  
heeman@cse.ogi.edu

**Géraldine Damnati**

France Télécom CNET DIH/RCP  
2 av. Pierre Marzin  
22307 Lannion Cedex, France  
damnati@lannion.cnet.fr

**Abstract** - Phrase-based language models have grown in popularity since they allow the speech recognition process to make use of more context in recognizing the words. Previous approaches have used perplexity reduction to identify groups of words to be linked into phrases and have used these phrases as the basis for computing the language model probabilities. In this paper, we argue that perplexity reduction is only one of three aspects to be considered in choosing the phrases. We also argue that the chosen phrases should not be the basis for computing the language model probabilities. Rather, the probabilities should be *derived* from a language model built at the lexical level.

## 1 Introduction

Most research in speech recognition is based on using lexical units as the interface between acoustic and language modeling. The acoustic model rates the likeliness of a stretch of sound given a certain word and the language model rates the likeliness of the word given the past words that were recognized. However, using lexical units as the coupling between the acoustic and language modeling is problematic.

1. During the first pass of a speech recognizer, one usually uses a bigram model due to efficiency concerns since this allows a smaller state space. However, trigram models can capture more of the prior context for predicting the next word. Hence, having acoustic units that encapsulate several ‘words’ will result in better probability estimates, since more context can be incorporated into the bigram.
2. The units that are appropriate for higher level analysis are not always realized acoustically as a distinct word. For instance in English, “can not” is usually contracted to “can’t”, and the word pair of “want to” is shortened to “wanna”. In French, this phenomena is much more widespread, especially for word pairs in which the second word begins with a vowel. *Elision* is where part of a word is cut with regards to pronunciation, especially for determiners, such as in “l’information”, and objective pronouns, as in “m’appelle”. *Liaison* is where

---

<sup>1</sup>Presented at the IEEE Workshop on Automatic Speech Recognition and Understanding, Santa Baraba CA, December 1997, ©IEEE.

the pronunciation of a word is changed because of its lexical context, as in “ils ont”, in which “ont” is pronounced with a beginning “z” sound. Acoustically, we should treat these as single units.

3. Even if no change in pronunciation occurs with a word pair, the first word might be difficult to recognize in isolation of the following word. This is especially true for function words, such as in the phrase “the weather” or in the French phrase “un serveur”, which tend to be very short in duration. Using larger acoustic units in which the function word is grouped together with a content word could help the speech recognizer better deal with such words.

In common practice, the more frequent contractions in English are added to the vocabulary and considered as acoustic units. More generally, several researchers have proposed using acoustic units larger than lexical words, which they refer to as *phrase-based* language models (e.g. [3, 6, 8, 7]). These works have focused on the first issue discussed above: using phrases to capture larger dependencies for predicting the next word during the bigram pass in speech recognition. They use measures such as perplexity reduction to identify phrasal units. Successively, choose the word pair  $w_i w_j$  such that rewriting all occurrences of  $w_i w_j$  as the single token  $w_i-w_j$  results in the largest decrease in perplexity. Once a set of phrases has been selected, the training corpus is rewritten with the phrases replaced by the single tokens. From this rewritten training corpus, the language model probabilities are computed.

The above approach of selecting the phrasal units and computing the language model probabilities suffers from a number of shortcomings. First, a selection criteria based on minimizing perplexity just addresses the first goal of using phrase-based language models. The selection criteria must also take into account pronunciation effects and the difficulty in recognizing words in isolation.

The second problem concerns how the resulting probabilities are computed. Rewriting the corpus leads to an increase in vocabulary size, while decreasing the size of the training data. Especially when dealing with small corpora, this will lead to poorer probability estimates. Furthermore, if one wants to use a class-based approach [1, 2], the acoustic units of the rewritten corpus are probably not meaningful to cluster. Likewise, if one uses POS tags [4], concatenating words together will result in a proliferation of complex POS tags, which will probably be undertrained. Thus, we have a divergence between the acoustic units that best serve the speech recognizer and the units appropriate for language modeling: acoustic modeling prefers larger units which cause undertraining of the language model.

The third problem is that the approach of rewriting the corpus makes the phrases into indivisible units. If “I want” is chosen as a phrase, all occurrences of “I want” are rewritten as the single token “I\_want”, and a phonetic entry is given for this token. The speech recognizer will only be able to recognize the sequence of “I” and “want” by the phrasal phonetic entry.<sup>2</sup> In spontaneous speech, speakers sometimes

---

<sup>2</sup>Actually, if the language model probabilities are computed by interpolating the bigram probabilities with the unigram probabilities [5], then there will be a small probability assigned to recognizing the phrase “I want” as two separate tokens due to the unigram probability. However, this probably will be much smaller than the probability of recognizing “I want” as a single token.

pause during the middle of a construction. If they happen to pause during the middle of a sequence of words that was chosen as a phrasal unit, then it will be extremely unlikely that the words in the phrase will be recognized, since the phonetic entry for the phrase undoubtedly does not allow a silence to be inserted in the middle of it. Hence, the speech recognizer should have both alternatives available to it.

In the rest of this paper, we propose a new approach to computing the probabilities for a phrase-based language model. In this approach, we derive the phrase probabilities from a language model built using the lexical units. We show that this approach does better than computing the probabilities from the rewritten corpus. Furthermore, this approach is not limited to phrases that are chosen using perplexity. In fact, we show that a further word error-rate reduction can be achieved by adding in more phrases, without regard to whether they have a corresponding perplexity reduction.

## 2 Deriving the Language Model Probabilities

Due to the problems given above with computing the language model probabilities from the rewritten corpus, we compute the probabilities by *deriving* them from a language model built using the lexical units. Since some of the acoustic units are composed of several lexical units, we derive these probabilities using not only the bigram distribution of the lexical units, but also the trigrams and 4-grams where appropriate. We define the derived bigram probability  $\text{Pr}_d(p|q)$  as follows, where  $\text{Pr}_2$ ,  $\text{Pr}_3$  and  $\text{Pr}_4$  are the bigram, trigram, and 4-gram probability distributions from the lexical language model.<sup>3</sup>

### Equation 1

$$\text{Pr}_d(p|q) = \begin{cases} \text{Pr}_2(p|q) & \text{if } p \text{ and } q \text{ are both lexical units} \\ \text{Pr}_3(p|q_1q_2) & \text{if } p \text{ is lexical and } q \text{ is the phrase } q_1q_2 \\ \text{Pr}_3(p_2|qp_1)\text{Pr}_2(p_1|q) & \text{if } p \text{ is the phrase } p_1p_2 \text{ and } q \text{ is lexical} \\ \text{Pr}_4(p_2|q_1q_2p_1)\text{Pr}_3(p_1|q_1q_2) & \text{if } p \text{ is the phrase } p_1p_2 \text{ and } q \text{ is } q_1q_2 \end{cases}$$

To illustrate the above equation, assume that “it’s” and “gonna” are two of the phrases that were chosen, where “it’s” is composed of the lexical units “it” and “is”, and “gonna” is composed of “going” and “to”. The phrasal probability of “gonna” given “it’s” is given by the following.

$$\text{Pr}_d(\text{gonna}|\text{it's}) = \text{Pr}_4(\text{to}|\text{it is going})\text{Pr}_3(\text{going}|\text{it is})$$

As can be seen above, we use the trigram and 4-gram probability distributions to compute the bigram probabilities of the phrases, and hence take advantage of using the richer context that trigrams and 4-grams provide.

In the experiments given in the next section, we contrast the effect of using the richer context afforded by using the bigram, trigram and 4-gram lexical probabilities

<sup>3</sup>We have currently only investigated phrases made up of at most two lexical units. The equation below can easily be extended to larger phrases.

with that of only using the bigram lexical probabilities. For the bigram case, we use the following approximation.

**Equation 2**

$$\Pr'_d(p|q) \approx \begin{cases} \Pr_2(p|q) & \text{if } p \text{ and } q \text{ are both lexical units} \\ \Pr_2(p|q_2) & \text{if } p \text{ is lexical and } q \text{ is the phrase } q_1q_2 \\ \Pr_2(p_2|p_1)\Pr_2(p_1|q) & \text{if } p \text{ is the phrase } p_1p_2 \text{ and } q \text{ is lexical} \\ \Pr_2(p_2|p_1)\Pr_2(p_1|q_2) & \text{if } p \text{ is the phrase } p_1p_2 \text{ and } q \text{ is } q_1q_2 \end{cases}$$

For the example given above, the derived probability of “gonna” given “it’s” would be as follow.

$$\Pr_d(\text{gonna}|\text{it's}) \approx \Pr_2(\text{to}|\text{going})\Pr_2(\text{going}|\text{is})$$

With this approximation, the derived phrase-based language model uses the same probabilities for a sequence of words as the bigram lexical model.

The first advantage of this approach is that we can use the word-based probabilities to compute the phrase probabilities, which we argued above is a more appropriate level for language modeling. Since larger order  $n$ -grams can be used in computing the phrase probabilities, as illustrated by Equation 1, we can still take advantage of the increase in context that the phrases afford.

The second advantage of deriving the probability distributions from the lexical units is that we can in fact add in any phrases if there is a reason to think that they can be better utilized by the speech recognizer, without having to worry about whether the new units are suitable for estimating the probabilities. Hence, we can throw in acoustic phrases that are appropriate for the second and third points we mentioned earlier.

The third advantage is that we keep the probabilities for the original lexical items. Consider the French phrase “je veux”, which means “I want”. As is illustrated in Figure 1, the speech recognizer has both alternatives available: the path consisting of the single acoustic unit “je veux” as well as the path consisting of the two separate acoustic units “je” and “veux”. The probability that “je” follows a given word  $w$

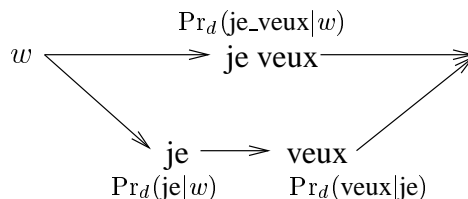


Figure 1: Flexible Phrase Recognition

does not change for the phrase-based approach. Rather, we add in extra probability corresponding to the phrases that “je” begins. The language model gives the same

preference to either derivation given in the figure, and simply lets acoustic factors determine which path is taken.<sup>4</sup> Hence, pauses between the words of the phrase can be modeled by taking the path in which the words are distinct acoustic units. Although the total probability mass for a given lexical context will now sum to more than one, this is only due to the duplication being added by the phrase entries.<sup>5</sup> In fact, one could view the alternative paths that can realize a sequence of words in much the same way that one views multiple phonetic entries for the same word. In estimating the probabilities, one doesn't take into account the number of phonetic variations. Phrases should be viewed the same way.

### 3 Experiments

Experiments were run on a French corpus of spontaneous human-computer spoken dialogues collected over a telephone network using the AGS voice service directory inquiry demonstrator [9]. Transcription of the dialogues has been carried out. The corpus was divided into a training set of 5451 speaker turns consisting of 26111 words and a test set of 802 speaker turns consisting of 4018 words. The vocabulary contains 874 different words, 76 of these words only occur in the test data. Speech recognition tests were carried out using a HMM-based, speaker independent, continuous speech recognizer.

Word error rate results for using phrases with the speech recognizer are given in Table 1. In this table, we compare two approaches: rewriting the corpus and deriving probabilities. The 35.08% reference is from a class-based bigram model computed with the lexical units (no phrases) from which the derived models are computed according to Equation 2. Hence, the results given in the third column do not take advantage of the extra context that the phrases provide. The first set of phrases was obtained automatically by a variant of the perplexity minimization method in which only words that occur at least 60 times are considered in constructing the phrases. This method leads to a small number of phrases (34). We can see in the first row that deriving probabilities outperforms both the word-based and the class-based bigram language model computed over the rewritten corpus. The other set of phrases consists of about 300 manually picked phrases, which we chose to correspond to the second and the third issue presented in Section 1. Here again, deriving outperforms rewriting. Furthermore, manually adding in phrases that are typically hard to acoustically recognize leads to a lower word error rate (32.72%). In regard to the results of the third column, the improvements are due to allowing the speech recognizer to use alternative models for phrases, as shown in Figure 1. It is also due to the fact that the vocabulary of the rewritten corpus has increased by 28%, while the size of the training corpus has decreased by 32%. Thus there is effectively less data to train models built from the rewritten corpus. As for the second column, it is interesting

---

<sup>4</sup>This is actually only true if we derive the phrase probabilities using just the bigram lexical probabilities. When deriving the phrase probabilities from the bigrams, trigrams and 4-grams of the lexical model, the phrasal probabilities of the phrase version (e.g. "I\_want") will make use of the richer context.

<sup>5</sup>We have found that normalizing the probabilities, or subtracting the phrase probabilities from the corresponding word probabilities, results in an increase in word error rate.

to notice that building a class-based language model from a rewritten corpus is not satisfactory, especially for the larger set of phrases. Again, this could be due to the reduced size of the training corpus leading to undertrained classes.

Selection	Rewritten Corpus		Derived Probabilities
	word	classes	
Perplexity Minimization	36.11%	36.03%	33.99%
Manually picked phrases	35.26%	37.82%	32.72%
No phrases	35.08%		

Table 1: Comparison of rewriting and deriving the probabilities

The results given in Table 1 did not make use of the full context in deriving the language model. Table 2 gives the results of the next set of experiments, in which we demonstrate that a further improved can be realized by deriving the phrase-based probability estimates using Equation 1, which uses the full context. Due to implementation concerns, we demonstrate this using the POS-based language model given in [4]. For this experiment, we use all bigrams that occur at least eight times as the set of phrases. The first row gives the baseline performance of the POS-based bigram model trained on the lexical units. The second row gives the results of using the bigram POS model of the first row to derive the probabilities of the phrase-based language model, using Equation 2. As in Table 1, we see a reduction in the word error rate, this time from 36.03 to 32.06. Hence, without changing the probabilities that the speech recognizer uses, but simply letting it deal with larger acoustic units results in an 11% reduction in the word error rate. The third row adds in the richer probability estimates that are afforded by using the lexical trigrams and 4-grams where appropriate to derive the phrase probabilities (using Equation 1). We see that this gives a further 4% reduction of in the word error rate. Together, we achieve a word error rate reduction using phrases of 14%. Thus, we see that most of the improvement is not due to using larger context, but simply letting the speech recognizer deal with larger acoustic units.

In Table 3, we investigate the types of phrases that meet the second and third criteria mentioned in the introduction (we use Equation 2 to derive the phrase-based model in order to factor out the effect of the first criteria). The first row gives the results from using all of the bigrams (511 of them) that occur at least eight times.

Model	Word Error Rate
No Phrases	36.03%
Derived from bigram model (Equation 2)	32.06%
Derived from 4-gram model (Equation 1)	30.94%

Table 2: Effect of longer order context (using POS model)

Model	Number of Phrases	Word Error Rate
All Phrases	511	32.06%
Syntactic Units	315	32.10%
Elisions	31	34.04%

Table 3: Effect of phrase selection

The second row removes bigrams that cross a major syntactic boundary, such as “serveur de”. The third row only includes bigrams involving *Elisions*, which would be similar to using contractions in English. From contrasting the first and second rows, we see that phrases that cross syntactic boundaries do not play a significant role for the second and third criteria. However, when just the syntactic units are used with the 4-gram model, the resulting word error rate is 31.27 in comparison to the rate of 30.94 achieved using all phrases. Hence, the phrases that are not syntactically motivated do play a role, but only with respect to the first criterion. The results of the third row show that it is not just the elision (contractions) that benefit from the improved acoustic modeling. However, in comparison to the baseline results of the POS-based word model given in Table 2 of 36.03, we see that the phrases due to elision account for roughly half of the improvement in the word error rate.

## 4 Conclusion

Previous work on phrase-based language models has been carried out using orthographically marked phrases (contractions in English) and phrases determined using perplexity reduction. This work has been *pre-language model*: the phrases are treated as indivisible units and used to rewrite the corpus, from which the language model probabilities are computed. In this paper, we propose a *post-language model* approach in which the phrase-based language model is derived from a model built at the lexical level (traditional techniques such as building class-based or POS-based models being held at the lexical level). We find that deriving the probabilities leads to reduced word error rates as it overcomes both the problem of treating phrases as indivisible units and rewriting the corpus, which leads to an undertrained language model. Our approach also shows that phrase selection should not be limited to perplexity reduction. Perplexity reduction only finds phrases that improve the context for word prediction and does not address acoustic considerations. Our work indicates that phrase-based language models need to address this concern. Work is still needed in order to automatically select the appropriate phrases.

## Acknowledgments

This research was carried out while the first author was visiting France Télécom CNET. The authors wish to thank Denis Jouvét, David Sadek, Alain Cozannet and Jacques Simonin for helpful comments.

## References

- [1] P. F. Brown, V. J. Della Pietra, P. V. deSouza, J. C. Lai, and R. L. Mercer. Class-based  $n$ -gram models of natural language. *Computational Linguistics*, 18(4), 1992.
- [2] G. Damnati and J. Simonin. A novel tree-based clustering algorithm for statistical language modeling. In *Proceedings of the 5<sup>th</sup> European Conference on Speech Communication and Technology*, 1997.
- [3] E. Giacchin, P. Baggia, and G. Micca. Language models for spontaneous speech recognition: a bootstrap method for learning phrase bigrams. In *Proceedings of the 3rd International Conference on Spoken Language Processing*, 1994.
- [4] P. A. Heeman and J. F. Allen. Incorporating POS tagging into language modeling. In *Proceedings of the 5<sup>th</sup> European Conference on Speech Communication and Technology*, 1997.
- [5] F. Jelinek and R. L. Mercer. Interpolated estimation of markov source parameters from sparse data. In *Proceedings, Workshop on Pattern Recognition in Practice*, pages 381–397, Amsterdam, 1980.
- [6] P. E. Keene, M. O’Kane, and H. G. Pearcy. Language modelling of spontaneous speech in a court context. In *Proceedings of the 4<sup>th</sup> European Conference on Speech Communication and Technology*, 1995.
- [7] G. Riccardi, A. L. Gorin, A. Ljolje, and M. Riley. A spoken language system for automated call routing. In *Proceedings of the International Conference on Audio, Speech and Signal Processing*, 1997.
- [8] K. Ries, F. D. Buø, and A. Weibel. Class phrase models for language modeling. In *Proceedings of the 4rd International Conference on Spoken Language Processing*, 1996.
- [9] M. D. Sadek, A. Ferrieux, A. Cozannet, P. Bretier, F. Panaget, and J. Simonin. Effective human-computer cooperative spoken dialogue: The AGS decomonstrator. In *Proceedings of the 4rd International Conference on Spoken Language Processing*, 1996.