ELSEVIER

Speech Communication 36 (2002) 81–95

SPEECH COMMUNICATION

www.elsevier.com/locate/specom

Integrated recognition of words and prosodic phrase boundaries

F. Gallwitz [*,1], H. Niemann, E. Nöth, V. Warnke [1]

University of Erlangen-Nuremberg, Chair for Pattern Recognition, Martensstr. 3, 91058 Erlangen, Germany

Received 3 February 2000; received in revised form 5 June 2000; accepted 24 November 2000

Abstract

In this paper, we present an integrated approach for recognizing both the word sequence and the syntactic–prosodic structure of a spontaneous utterance. The approach aims at improving the performance of the understanding component of speech understanding systems by exploiting not only acoustic–phonetic and syntactic information, but also prosodic information directly within the speech recognition process. Whereas spoken utterances are typically modelled as unstructured word sequences in the speech recognizer, our approach includes phrase boundary information in the language model and provides HMMs to model the acoustic and prosodic characteristics of phrase boundaries. This methodology has two major advantages compared to purely word-based speech recognizers. First, additional syntactic–prosodic boundaries are determined by the speech recognizer which facilitates parsing and resolve syntactic and semantic ambiguities. Second – after having removed the boundary information from the result of the recognizer – the integrated model yields a 4% relative word error rate (WER) reduction compared to a traditional word recognizer. The boundary classification performance is equal to that of a separate prosodic classifier operating on the word recognizer output, thus making a separate classifier unnecessary for this task and saving the computation time involved. Compared to the baseline word recognizer, the integrated word-and-boundary recognizer does not involve any computational overhead.

Zusammenfassung

In diesem Artikel stellen wir einen integrierten Ansatz zur Erkennung der Wortkette und der syntaktisch–prosodischen Struktur einer spontansprachlichen Äußerung vor. Ziel des Ansatzes ist es, die Leistungsfähigkeit von sprachverstehenden Systemen dadurch zu verbessern, dass nicht nur akustisch–phonetische und syntaktische Information, sondern auch prosodische Information direkt im Rahmen des Spracherkennungsprozesses genutzt wird. Üblicherweise werden gesprochene Äußerungen innerhalb des Spracherkenners als unstrukturierte Wortfolgen modelliert. In unserem Ansatz werden Phrasengrenzen dagegen direkt in das Sprachmodell integriert, und HMMs werden zur Modellierung der akustischen und prosodischen Eigenschaften von Phrasengrenzen herangezogen. Hierdurch ergeben sich zwei wesentliche Vorteile gegenüber rein wortbasierten Spracherkennern: Zum einen werden zusätzliche syntaktisch–prosodische Grenzen durch den Spracherkenner bestimmt, die von einem nachgeschalteten Parser zur Beschleunigung und Disambiguierung genutzt werden können. Zum anderen konnte – auch ohne Berücksichtigung der erkannten Grenzen – mit dem integrierten Ansatz eine relative Verbesserung der Wortfehlerrate um 4% erzielt werden,

* Corresponding author.
E-mail address: gallwitz@sympalog.de (F. Gallwitz).
[1] Now with Sympalog Speech Technologies AG, Erlangen.

0167-6393/02/$ - see front matter
PII: S0167-6393(01)00027-9

verglichen mit dem traditionellen wortbasierten Ansatz. Die Güte der Grenzklassifikation entspricht dabei der eines separaten prosodischen Klassifikators, der auf dem Worterkennungsergebnis aufsetzt; ein solcher wird also für diese Aufgabe nicht mehr benötigt und die hierfür benötigte Rechenzeit wird eingespart. Verglichen mit dem reinen Worterkenner benötigt der integrierte Erkenner für Wörter und Grenzen keinerlei zusätzliche Rechenzeit.

Keywords: Speech recognition; Prosody; Speech understanding

1. Introduction

Today, there appears to be a general consensus in the speech recognition community that the area of speech recognition is only concerned with the problem of finding the *sequence of words* associated to a given acoustic observation. Accordingly, the term 'speech recognition' is usually defined in the following manner (similar definitions can be found, for example, in (Jelinek, 1997; Schukat-Talamazzini, 1995):

> The automatic speech recognition problem consists of finding the sequence of words $\vec{W}$ associated to a given acoustic sequence $\vec{X}$. (Beccetti and Ricotti, 1999, p. 8)

It is well known, however, that the word sequence associated to an utterance does not always contain all the information that is necessary for *understanding* its meaning, because an important source of information is usually not captured by the word sequence: *prosody*. The problem of classifying prosodic phenomena, such as phrase boundaries, sentence mood and accentuation, has therefore received a lot of attention in recent years. As a result, the first speech understanding systems have emerged that take into account prosodic information (Kompe, 1997). The classification of prosodic phenomena, however, has typically been regarded as a task which could be treated either independently from the problem of recognizing the spoken word sequence, or as a subsequent step.

In this paper, we investigate an integrated approach that combines the recognition of the spoken word sequence (i.e. speech recognition in the above sense) and the classification of prosodic phrase boundaries in a single search procedure. Instead of regarding speech as an unstructured sequence of words, speech is modelled as a sequence of words and phrase boundaries. The resulting recognizer for words and prosodic phrase boundaries is still a *speech recognizer* according to the following, more general, definition:

> Speech recognition can be generally defined as the process of transforming a continuous speech signal into discrete representations which may be assigned proper meanings and which, when comprehended, may be used to affect responsive behaviour. (Lea, 1980b, p. 40)

In spoken language, especially in spontaneous speech, prosodic boundaries are of similar importance for understanding an utterance as punctuation marks are in written language. Words which 'belong together' from the point of view of meaning are grouped into *prosodic phrases*, and it is widely agreed upon that there is a high correspondence between prosodic and syntactic phrase boundaries (Wightman et al., 1992; Kompe, 1997).

Prosodic boundaries are often marked by silence periods, and sometimes by filled pauses, such as 'uh', and they are usually indicated by specific energy and fundamental frequency (F_0) contours and by durational variations of the surrounding syllables (Kießling, 1997). Also, as punctuation marks in written language, they are often predictable from the surrounding word context.

In automatic speech understanding, this information may be important even in the context of a comparatively simple application, such as an automatic train timetable information system. Consider, for example, the following user utterances:

U1: *Of course not on Monday.*

U2: *Of course not. On Monday*!

The question whether a prosodic phrase boundary occurred after the word '*not*' is crucial

for the semantic interpretation of the word sequence and for determining the next system utterance. Depending on the phrasing, one of the following two utterances may be appropriate:

S1: *What day would you like to travel*?

S2: *You would like to travel on Monday*?

Selecting the wrong response (**S1** for **U2**, or **S2** for **U1**) will most certainly annoy the caller and will probably make her/him hang-up.

It might be argued that the correct interpretation of the word sequence could also be determined without prosodic information, if the dialogue history is taken into account. Depending on the previous system utterance, at least one of the two above interpretations could be declared illogical. This involves a considerable amount of higher-level knowledge and "intelligent" processing, however, whereas prosodic information in the speech signal can directly resolve the ambiguity. Furthermore, there is no reason to ignore information that may without a doubt contribute to finding the correct semantic interpretation, even if a sufficiently intelligent dialogue module is available (Kompe, 1997, Section 8.4).

The first speech understanding system to really integrate prosodic information into the understanding process is the German VERBMOBIL speech-to-speech translation system for appointment scheduling dialogues (Wahlster, 1993; Bub and Schwinn, 1996). In the VERBMOBIL prototype, prosodic information is calculated on the basis of the speech signal and the word recognition result. This information is used in various system modules, mainly for resolving syntactic and semantic ambiguities, and has been shown to significantly improve the total system performance (Kompe, 1997). For example, VERBMOBIL is able to provide different English translations for German utterances that contain the same word sequence but are prosodically distinct (Kompe, 1997):

Ja zur Not geht's auch am Samstag.
(*Well, if necessary, Saturday is also possible.*)
Ja. Zur Not. Geht's auch am Samstag?
(*Okay. If necessary. Is Saturday possible as well*?)

Speech recognition and prosodic analysis are performed in two separate modules, however, and the speech recognizer itself is only concerned with finding an optimal sequence of words (or a *word graph*, a graph of competing word hypotheses) that covers the whole speech signal. That is, the prosodic and the syntactic structures of the utterance are neither determined nor taken into account by the speech recognizer. The basic structure of the word recognition and prosody classification modules in VERBMOBIL is depicted in Fig. 1(a). A similar sequential architecture where prosody is classified on the basis of the word recognition result has also been used by other groups, e.g. (Stolcke et al., 1998).

We believe that syntactic–prosodic boundary information is also useful in an earlier stage of spontaneous speech processing. State of the art speech recognizers are typically based on two sources of knowledge: acoustic information and language model information. Statistical language models as used in most speech recognizers provide the probability of a given word sequence based on a rather simple model: it is assumed that a spoken utterance is an unstructured sequence $w_1, w_2, \ldots, w_n$ of words. This is not the case, however. It is intuitively clear that words at the beginning of a new phrase correlate less strongly with the last word of the preceding phrase than words within the same phrase. This can easily be demonstrated empirically. On VERBMOBIL utterances from a sub-corpus which is not part of the training corpus, for example, the amount of unseen word pairs is almost three times as high for word pairs *across* phrase boundaries than for word pairs within phrases. Whereas only 14% of the word pairs within phrases have not been observed in the training corpus, the same ratio for words pairs across phrase boundaries is 38%. Any n-gram language model will provide lower probabilities for word transitions that have not been observed in the training data. That is, language model probabilities across phrase boundaries are *systematically underestimated* by traditional, word-based language models.

A similar effect has also been found in the neighbourhood of filled pauses (Shriberg and Stolcke, 1996). As a consequence, a language model for spontaneous speech is proposed in (Stolcke and Shriberg, 1996), where different types of disfluencies (filled pauses, repetitions and deletions) are predicted, and probabilities of following

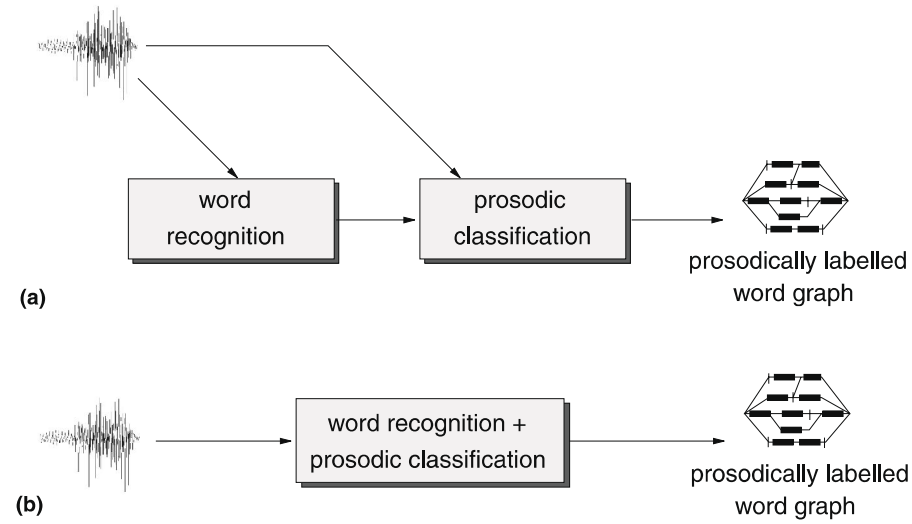

Fig. 1. The sequential approach to word recognition and prosody classification which has been successfully applied in the VERBMOBIL speech-to-speech translation system (Kompe, 1997; Kießling, 1997) and has also been used by other groups, e.g. (Stolcke et al., 1998), and the integrated approach proposed in this paper. (a) Sequential approach, (b) integrated approach.

words are estimated on the basis of the fluent word sequence that was supposedly intended by the speaker. This approach, however, did not have a significant impact on the recognition accuracy. One of the reasons for this result is noted in (Stolcke and Shriberg, 1996): phrase (or clause) boundaries grossly violate the assumptions of the proposed model, because filled pauses strongly correlate with boundaries of linguistic segments. Thus, 'cleaning up' the surrounding words to remove the disfluency can be counterproductive.

In our integrated approach which is depicted in Fig. 1(b), phrase boundaries are directly integrated into the language model, and silent and filled pauses are allowed to occur in two different functions: either they are syntactically insignificant and thus ignored in the language model ('clean-up'), or they occur at phrase boundaries. These two different functions of pauses have been described earlier in (O'Shaughnessy, 1992), where the first type of pause was referred to as *ungrammatical* (or *unintentional*), and the second type as *grammatical* (or *intentional*). Although a more detailed discrimination of different pause functions may be possible, we found this two-class model especially suitable for an integration into the recognizer search procedure.

Furthermore, in our model, phrase boundaries are also allowed to occur at fluently spoken word–word transitions. The fact that a word is separated from its predecessor by a phrase boundary should contribute a great amount of information when language model probabilities are calculated, while the preceding word is less significant. By integrating models for syntactic–prosodic phrase boundaries into the word recognizer and into the statistical language model, the word recognizer can incorporate information about the structure of the utterance. An integrated model of sequences of words and boundaries allows for a distinction between word transitions across phrase boundaries and transitions within a phrase, which is an obvious advantage.

An entertaining but representative example that clearly shows the advantages of an *integrated* processing of word information and prosodic information as proposed in this paper is given in (Lea, 1980a, p. 167):

A: *What is that in the road ahead*?

B: *What is that in the road*? *A head*?

Here, not just the semantic interpretation, but also the *word sequence* depends on the prosodic structure of the utterance. That is, if prosodic information is taken into account in this example, it will be considerably more helpful if it is integrated into the word recognition process.

In phrase boundary recognition experiments based on word recognizer results, it has been shown that prosodic features can significantly improve the detection accuracy of syntactic phrase boundaries compared to a pure language model-based approach (Kompe, 1997). This is especially the case with syntactically ambiguous boundaries, as in the above example utterances. In this paper, we also investigate how additional prosodic information can be incorporated into our integrated approach to recognize words and syntactic–prosodic boundaries.

As the feature set used for our separate boundary classifier is not suitable for the system architecture of the integrated word-and-boundary recognizer, we developed new frame-based prosodic feature sets that incorporate information on the fundamental frequency and energy contours as well as durational information.These features are used as input to an ANN in order to calculate the prosodic probability of a phrase boundary for each time frame. The resulting probabilities are then utilized as a second input stream to the HMM based recognizer, in addition to the acoustic–phonetic probabilities that are based on a cepstral feature vector and a Gaussian codebook. Thus, the integrated recognizer combines three sources of information: acoustic–phonetic information, prosodic information, and language model information.

The research presented in this paper is described in more detail in (Gallwitz, 2001). This thesis also addresses other problems of spontaneous speech recognition which are not discussed in this paper, such as the integrated detection and classification of out-of-vocabulary words.

The remainder of this paper is structured as follows. In Section 2, we review some important publications which are related to the work described this paper. In Section 3, we briefly describe the phrase boundary labelling system that was used as a basis of our experiments. In Section 4, the treatment of phrase boundaries during training and recognition in our approach is described. In Section 5, a hybrid HMM–MLP system architecture is presented that incorporates prosodic features into the recognition process. The prosodic feature sets employed in our experiments are described in Section 6. The training procedure of the hybrid speech recognizer is then discussed in Section 7. Finally, experimental results are given in Section 8. The paper closes with a brief summary of the main results.

2. Related work

To our knowledge, no approach has been published earlier where syntactic–prosodic structure was directly integrated into the speech recognition process. A number of studies have been performed, however, where information about the syntactic–prosodic structure of utterances was used for a rescoring of the n-best sentence hypotheses, or for a rescoring of word graphs.

In (Veilleux and Ostendorf, 1993; Ostendorf, 1994), two models which predict prosodic phrase boundaries and accents were employed for rescoring the n-best sentence hypotheses. The first model computes probabilities for prosodic classes with classification trees on the basis of a set of syntactic features which are determined from an automatic parse of the word chain. The second model computes similar probabilities from acoustic–prosodic features which are derived from the speech signal. The probabilities computed by both models were then combined in order to obtain a prosodic score for each sentence hypothesis. This score was then combined with the acoustic score and the (word-based) n-gram score computed by the word recognizer, and a rescoring of the n-best sentence hypotheses was performed. On a subset of the ATIS corpus, the average rank of the correct sentence hypothesis was improved by 23%. Although this approach is somewhat comparable to the approach presented here – it is also based on a combination of acoustic, prosodic and language model scores – it is not suitable for an integrated search procedure. This is due to the fact that the approach is based on syntactic features which

require the complete word chain to be parsed before the prosody model can be applied. More recently, an *n*-best rescoring approach was proposed in (Stolcke et al., 1999) which involved a 'hidden event' language model and also incorporated prosodic cues. With this approach, a slight word error rate (WER) reduction was achieved on the Switchboard corpus.

A considerably more complex integrated language model for spontaneous speech which incorporates speech repairs, phrase boundaries and discourse markers was proposed in (Heeman and Allen, 1999), where it was used for the classification of a number of different event classes on the basis of the spoken word sequence. A similar model was also employed for rescoring word graphs in (Heeman, 1999), which yielded a slight but significant WER reduction. The author did not comment on the computational requirements of his approach. These can be assumed to be rather high, due to the huge search space involved in extracting the optimal word sequence from a large word graph with his model. Different from the more simplistic integrated language model used in our approach which we first proposed in (Gallwitz et al., 1998), the model is not suitable for a direct integration into HMM-based speech recognizers, because it incorporates information which is not available during the left-to-right decoding procedure, such as part-of-speech tags for each word.

A word graph rescoring procedure for spontaneous speech which not only incorporates acoustic and syntactic information, but also prosodic information was presented by our group in (Warnke et al., 1999). The integrated search procedure involves the classification of different prosodic event classes as well as a dialogue act segmentation and classification. Because of this implicit (rough) semantic interpretation, we called the output of this procedure an *interpretation graph*. The approach was designed to optionally use word graphs produced by our integrated word-and-boundary recognizer as an input, which considerably reduces the search space involved compared to rescoring prosodically unmarked word graphs.

A sequential approach for the recognition of read, continuous Japanese speech with the help of information about the prosodic structure of an utterance was proposed in (Iwano and Hirose, 1999). In this approach, a first recognition pass was performed without prosodic information. Prosodic information and the result of the first recognition pass were then employed for the detection of prosodic boundaries. The utterance was then segmented according to the detected boundaries, and a second recognition pass was conducted with a specialized bigram language model for the resulting segments. As a result, a slight improvement in recognition rate was observed on a small test sample.

n-Gram language models which include both words and phrase boundary symbols comparable to the integrated word-and-boundary language model used in our approach have been used earlier for the purpose of phrase boundary classification on the basis of the word recognizer output, e.g. Kompe et al. (1995), Kompe (1997), Stolcke et al. (1998). Phrase boundary classification results on the basis of the recognized word sequence were published within this research in (Gallwitz et al., 1998), where they were used as a baseline for the proposed integrated approach and, independently, in (Stolcke et al., 1998). Here, a detailed evaluation was presented that demonstrated the considerable drop in classification performance when moving from the spoken word sequence as input to the prosodic classifier to an automatically recognized word sequence.

3. Syntactic–prosodic boundaries

The labelling scheme used for our experiments was originally designed for the purpose of improving the syntactic and semantic analysis of word graphs with the help of prosodic information. Starting point for the annotation of our material with syntactic–prosodic labels was the assumption that there is a strong – albeit not perfect – correlation between syntactic phrasing and prosodic phrasing (cf. Lea, 1980a; Vaissière, 1988; Price et al., 1991). This assumption could be corroborated earlier in experiments with German read speech where similar labels could be used successfully for the training of prosodic classifiers, cf. (Kompe et al., 1994). In order to save time, we

annotated these boundaries only using the written word chain. The 'syntactic–prosodic' boundaries relevant for our present purpose – we called them M3-boundaries – are those syntactic boundaries that are expected to be marked prosodically, as can be seen in the following example:

perhaps I should first introduce myself M3 *my name is Lerch*

Most M3 boundaries are labelled at boundaries between main clauses or subordinate clauses, at the boundaries of left and right dislocations and at the boundaries of embedded sentences/phrases. In the VERBMOBIL data, the average length of a prosodic phrase between two M3-labels is 7.1 words, while the average turn length is 22 words. Details on the data used in our experiments are given in Section 8. More details on our labelling scheme can be found in (Batliner et al., 1998).

4. Basic approach

The basic idea behind our approach is that phrase boundaries should be treated in the language model (LM) in a similar fashion as words. Thus, we provide a language model category (or word class) for phrase boundaries in the n-gram LM, and we provide HMMs to model the acoustic and prosodic characteristics of phrase boundaries.

In (Kompe, 1997), it has been shown that the syntactic–prosodic boundaries often happen to occur in combination with non-verbal noises, pauses or filled pauses. This makes it desirable to exploit the information that is provided regarding the correlation between boundaries on the one hand and pauses, filled pauses and non-verbals (NV) on the other hand. Thus, we incorporate this information by training suitable LM probabilities for different non-verbal phenomena which occur at phrase boundaries.

Furthermore, we assume that silence periods and non-verbals within phrases and across phrase boundaries can be discriminated based on acoustic and prosodic information. Thus, different HMMs are trained for pauses and non-verbals within phrases, and across phrase boundaries. For example, two different HMMs are used for the two occurrences of the filled pause '*uh*' in the utterance:

from Munich uh I want to travel on uh Saturday.

The first '*uh*' is modelled by the specific boundary model named '[M3-uh]', which is trained on all occurrences of '*uh*' at phrase boundaries, and the second '*uh*' is trained on all occurrences of '*uh*' within phrases.

We train HMMs for several combinations of boundaries and non-verbals, and include them in the statistical language model according to their syntactic function: Non-boundary models for pauses and non-verbals are skipped in the language model and boundary models are treated like words, both during training and decoding.

A special situation arises in the case of a phrase boundary that does not coincide with a filled pause or a non-verbal. Here, we provide a one-state HMM that always consumes one time frame. By consuming one time frame, the recognizer can incorporate information on the acoustic and prosodic characteristics of this type of phrase boundary. One 10 ms time frame, on the other hand, is too short to severely impair the recognition of the two neighbouring words. Another possible approach which we did not yet evaluate would allow for the optional introduction of phrase boundary symbols at word boundaries *without* consuming a time frame. This solution would only require some minor modifications of the recognizer search procedure. It is not clear, however, how prosodic information can then be efficiently incorporated.

During the word recognizer search procedure, several different situations have to be taken into account at transitions from word w_i to word w_{i+1}. The word recognizer implicitly makes a decision for the most probable alternative, based on the language model scores and on the acoustic and prosodic scores of the word-and-boundary HMMs involved (in the following, we only consider the bigram scores; the higher-order language model scores are calculated accordingly):

1. If no boundary or non-verbal is hypothesized, the bigram score $P(w_{i+1} \mid w_i)$ is used.

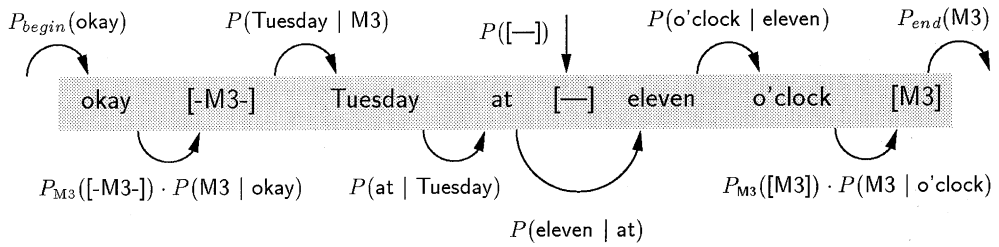

Fig. 2. The integrated word-and-boundary language model (in the case of a bigram-based recognizer) illustrated with the example utterance '*okay – Tuesday at – eleven o'clock*' (the dashes indicate silent pauses). The correct sequence of word-and-boundary models and the corresponding bigram probabilities are given in the figure. All M3 boundary models (e.g. [-M3-] for a boundary which is marked by a silent pause and [M3], which consumes only one time frame) are in a single language model category M3; the category-dependent emission probabilities for M3 models are denoted $P_{M3}(\cdot)$.

2. If an M3 boundary is hypothesized (possibly represented by an M3 silence model or an M3 non-verbal model), the bigram scores $P(\mathrm{M3}\,|\,w_i)$ when entering the M3-model and $P(w_{i+1}\,|\,\mathrm{M3})$ when entering w_{i+1} are employed. The scores are the transition probabilities for entering the *language model category* M3. When categories (or word classes) are used in a stochastic n-gram model, the LM score can be factorized in a transition probability and a category-dependent emission probability (the maximum-likelihood estimate of the emission probability for a particular word is given by the number of observations of this word divided by the total number of observations of words belonging to its category). In our case, the information that is provided regarding the correlation between certain types of pauses/non-verbals and M3 boundaries can be exploited by including all relevant models into a single language model category M3 and by using suitable category dependent emission probabilities $P_{\mathrm{M3}}(\cdot)$ for different non-verbal phenomena which occur at phrase boundaries.
3. If no boundary, but a non-verbal (NV) or silence period is hypothesized, the constant unigram probability $P(\mathrm{NV})$ is used when entering the NV model, and $P(w_{i+1}\,|\,w_i)$ when entering w_{i+1}. Thus, non-verbals or silence periods that do not mark syntactic boundaries are treated as random events that do not depend on the surrounding word context. Consequently, they are ignored when the probability of the following word is calculated.

In Fig. 2, the integrated word-and-boundary language model is illustrated with an example utterance.

The search algorithm of the recognizer (e.g. beam-search or $A^\star$ search) will now determine the optimal sequence containing words and boundaries. Alternatively, a word graph can be generated which contains additional phrase boundary labels.

5. System architecture

The proposed approach can be used with any state-of-the-art HMM-based speech recognizer, irrespective of the specifics of the HMM topology, the type of density, or the decoding algorithm. In particular, it can also be used within single-pass recognizer architectures. Only some slight modifications to the decoding algorithm might be necessary, to allow for the treatment of syntactically irrelevant silence-periods and non-verbals as described above. Even without additional prosodic information, the integration of phrase boundaries into the recognition process has been shown to yield improved word accuracies for spontaneous speech recognition (Gallwitz et al., 1998).

It is our goal, however, to incorporate additional prosodic information into the approach. Prosodic information, e.g. movements of the F_0 contour, can help to improve the detection of phrase boundaries, which implicitly – via the statistical LM – might also improve the word accuracy. Furthermore, ambiguous boundaries can only be reliably classified if prosodic information is

taken into account, which is especially important when the occurrence of a prosodic boundary has an impact on the semantic interpretation of an utterance.

In preliminary experiments, a direct integration of prosodic features into the feature vector used by the word recognizer did not yield any improvement. Instead, there was even a significant decline in word accuracy. The probable reason for this lies in the complex distributional properties of features that are derived from prosodic parameters, such as the second derivative of the F_0, which could not be accurately modelled by the Gaussian distributions employed in our word recognizer.

We have therefore developed a hybrid architecture that independently processes acoustic–phonetic and prosodic information on a level close to the signal. Both streams of information are then combined during the recognition process. Acoustic–phonetic information (i.e. mel–cepstral coefficients and their first derivatives) are processed as in our baseline recognizer, which is based on semi-continuous hidden Markov models (SCHMMs). This involves a soft vector quantization on the basis of a Gaussian codebook. Acoustic–prosodic features are used as input to a multilayer perceptron (MLP), which estimates the probability of a prosodic boundary in the current frame. The

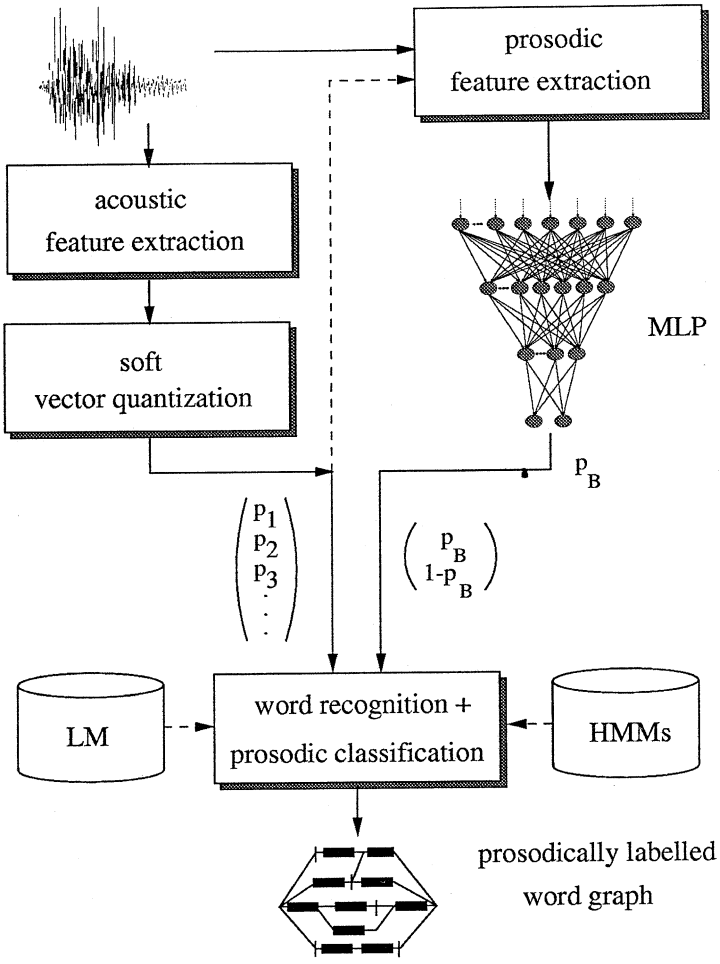

Fig. 3. Proposed architecture of an MLP–HMM hybrid system for integrated classification of prosodic boundaries using additional prosodic features.

architecture is depicted in Fig. 3. The dashed arrow indicates that information extracted from the stream of vector quantization results, e.g. durational information, may be included in the prosodic feature vectors. The two input streams of the word recognizer are treated as stochastically independent during the calculation of the HMM probabilities. A prosodic weight factor (similar to the linguistic weight factor for LM probabilities) is introduced to allow for a balancing of acoustic and prosodic information.

6. Prosodic features

The following acoustic parameters are considered to be the most valuable for the classification of prosodic information in ASU (Kießling, 1997, p. 67):

- energy (the acoustic correlate of loudness),
- the fundamental frequency F_0 (the acoustic correlate of pitch),
- pause-length, and
- phone duration.

Although there are obviously strong interdependencies between acoustic–phonetic and acoustic–prosodic information, we find it helpful to use the terms *acoustic–phonetic feature* and *acoustic–prosodic feature*. The purpose of acoustic–phonetic features is mainly to incorporate segmental, phonetic information over a short period of time; typically this time is in the order of the mean phoneme duration (about 70 ms). Acoustic–prosodic features cover suprasegmental information that is generally included in significantly larger portions of the speech signal.

As mentioned above, in the VERBMOBIL system, prosodic features are calculated based on a time alignment of the word recognition result (Kompe, 1997). This approach is now commonly used, because it allows for the incorporation of information about the position of word and syllable boundaries, and for a normalization of the features based on word, syllable or phoneme information. Unfortunately, this type of feature is not suitable for the integrated approach of recognizing words and prosodic information in one step, for the simple reason that no recognition result can be available before the recognition process even started. Instead, an incremental calculation of features should be possible without having to wait for the end of an utterance. Furthermore, all prosodic features have to be calculated frame-based, and only based on the speech signal (or on information that can be derived from the speech signal efficiently and incrementally, such as the vector quantization result). Thus, we developed a number of frame-based suprasegmental features which incorporate specific movements of prosodic parameters. These may indicate the occurrence of prosodic events. For example, it is worth looking at, whether the second derivative of the fundamental frequency, calculated over a fixed-sized window of 1 or 2 s, gives hints on the position of prosodic boundaries.

A part of the prosodic feature set used for our experiments which is based on F_0 values is shown in Table 1. All features are based on F_0 values that

Table 1
F_0-based suprasegmental feature set

F_1	$F_0(t)$	F_0 at time t
F_2	$\Delta_t^{10}(F_0)$	Delta coefficient at t with a context of ± 10 frames
F_3	$\Delta\Delta_t^{10}(F_0)$	Delta coefficient at t with a context of ± 10 frames
F_4	$\Delta_t^{20}(F_0)$	Delta coefficient at t with a context of ± 20 frames
F_5	$\mathrm{MSE}_t^{20}(F_0)$	Mean square error of F_0 and F_0 regression line within a 40 frame context centred at t
F_6	$\Delta\Delta_t^{20}(F_0)$	Delta delta coefficient at t with a context of ± 20 frames
F_7	$\Delta_t^{40}(F_0)$	Delta coefficient at t with a context of ± 40 frames
F_8	$\mathrm{MSE}_t^{40}(F_0)$	Mean square error of F_0 and F_0 regression line within a 80 frame context centred at t
F_9	$\Delta\Delta_t^{40}(F_0)$	Delta delta coefficient at t with a context of ± 40 frames
F_{10}	$\Delta_t^{80}(F_0)$	Delta coefficient at t with a context of ± 80 frames
F_{11}	$\mathrm{MSE}_t^{80}(F_0)$	Mean square error of F_0 and F_0 regression line within a 160 frame context centred at t
F_{12}	$\Delta\Delta_t^{80}(F_0)$	Delta delta coefficient at t with a context of ± 80 frames

were first transformed on a logarithmic scale and then linearly interpolated in unvoiced parts of the signal (the frame length is 10 ms). We employed a set of 24 features which also included features calculated on the basis of the energy contour, in a similar fashion. As yet, no experiments have been performed which explicitly include durational information. We are currently developing methods for extracting durational information from the result of the soft vector quantization. Obviously, this stream of symbols (with corresponding probabilities) can be used to detect lengthenings and variations in the speaking rate. For example, the amount of variation in the stream of the best-scoring codebook classes typically decreases in the presence of lengthenings.

7. MLP and HMM training

It is not straightforward to define the optimal output of the MLP in the hybrid architecture described above. Ideally, it should provide the phrase boundary probability 1.0 for frames that are associated with a boundary HMM, and 0 for non-boundary frames. This is not feasible, however, because prosodic boundaries cannot realistically be associated to one single time frame. Instead, indications for a prosodic boundary should also be expected in the surrounding frames. Restricting the MLP training set for the phrase boundary class to the comparatively small set of time frames which are directly associated to a boundary HMM is certainly sub-optimal.

Our solution to this problem is to define a heuristic goal function for the MLP which is based on the cosine function, as depicted in Fig. 4. Each peak corresponds to the first frame of a boundary HMM, when a forced alignment of the transliteration is performed. This goal function is used for training the MLP. During HMM training, the output of the trained MLP is used. In the worst case, the MLP output is not correlated with the occurrence of boundaries. This should not degrade the recognition performance compared to a system without prosodic information, however, because the resulting HMM emission probabilities for the boundary and non-boundary classes would be close to 0.5 for all HMM states, which means that the MLP output has no impact on the recognition result. Nevertheless, any correlation of the MLP output with phrase boundaries should improve the recognition results for phrase boundaries, because the HMM emission probabilities are then trained accordingly. The recognition of word HMMs is only affected if certain words typically occur in the neighbourhood of phrase boundaries. This effect is expected to have a positive influence on the word recognition performance.

This goal function was also used for evaluating the performance of the hybrid word-and-boundary recognizer in the case of an optimal performance of the MLP boundary classifier (see below). In this case, the ideal MLP output is used both during training and recognition. The latter, of course, is

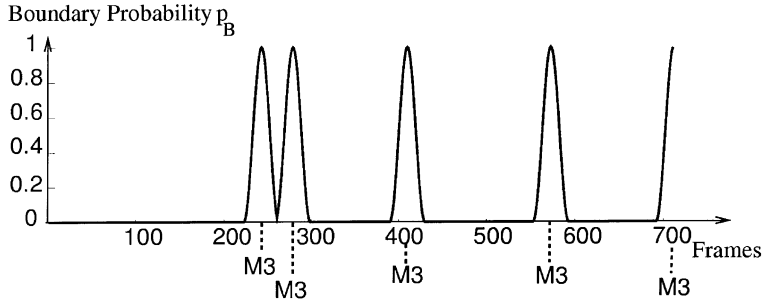

Fig. 4. The desired output of the MLP classifier, which was used for training the MLPs, and for the experiments involving 'ideal' boundary classification performance.

only possible if the position of phrase boundaries is available for the test sample.

8. Experiments and results

The experiments reported in this paper have been performed on a subset of the German VERBMOBIL corpus. The training, validation and test samples are shown in Table 2 (the figures for phrase boundaries do not contain the trivial boundaries at the beginning or end of a turn).

We used a speaker independent SCHMM word recognizer with a codebook size of 512 classes. No speaker adaptation was performed and only intra-word subword models (polyphones) were used. The 24D acoustic–phonetic feature set consists of 12 mel–cepstral features and their first derivatives. The recognizer architecture employed for our experiments uses a bigram LM in the first pass of the recognition process and a 4-gram LM in the second pass. The only difference between both language models is the context size involved. That is, the 4-gram LM can also be used in a single recognition pass when a suitable recognizer architecture for higher-order language models is available. The vocabulary size is 3166 words.

Six additional boundary models were used in the experiments involving phrase boundaries. These models are of similar structure as the corresponding models for pauses, filled pauses and non-verbals that are used for modelling non-verbal phenomena which do not coincide with phrase boundaries. These models consist of linear HMMs with three to nine HMM states which are trained in a partially unsupervised training procedure. For example, the filled pause '*um*' is represented by six HMM states; three for each phoneme. The models are initialized on the basis of the pauses and non-verbal phenomena labelled in the transliteration. During HMM training, at each word boundary where no phrase boundary is labelled, any HMM from the set of non-verbal HMMs may be optionally inserted during the forced Viterbi alignment. Similarly, a HMM from the set of boundary HMMs is automatically selected from the set of boundary HMMs at word boundaries where a phrase boundary is labelled. That is, the training procedure of the integrated approach only requires the labelling of non-verbal phenomena for the initialization step, which can also be performed on a small subset of the training data. For the training procedure, only the sequence of words and phrase boundaries is required, whereas the training of the baseline word recognizer's language model requires a detailed transliteration which includes pauses, filled pauses and non-verbal phenomena.

Table 2
Training, validation and test data

Sample	Turns	Words	M3 phrase boundaries
Training	11.714	258.956	24.382
Validation	48	1044	89
Test	268	4783	500

The word accuracies are calculated based on the pure word chain, i.e. the boundary labels were removed from the recognizer result in the case of the integrated approach. The evaluation of the recognized boundaries is performed in the following manner. First, an alignment based on the minimum Levenshtein-distance criterion is performed between the recognized word chain and the reference transliteration. During this procedure, the boundary labels are treated just like words. A similar but slightly different alignment procedure was used for the evaluation of boundary classification performance on the basis of recognized word chains in (Stolcke et al., 1998), where additional labels were included for fluent word transitions when the alignment was conducted.

After the alignment is performed, all pairs of hypothesized symbols and reference symbols that include at least one boundary are used to calculate precision and recall rates. Note that even a perfect boundary classification will not result in 100% precision and recall if a certain number of word recognition errors are present in the recognition result, because these can lead to a mismatch between the alignment of the reference and the hypothesized word-and-boundary sequence.

The recognition results are given in Table 3. The results were evaluated with respect to the WERs and recall (RCL) and precision (PRC) rates for M3 phrase boundaries. The corresponding real

Table 3
Word recognition and boundary classification results

System	WER	RCL	PRC	RTF
Baseline word recognizer	23.8%	–	–	4.1
Prosodic classifier	–	75.1%	74.7%	0.4
Integrated w&b recognizer	22.9%	74.5%	75.7%	4.0
Hybrid w&b recognizer	22.9%	75.7%	75.3%	4.2
Hybrid with 'ideal' MLP	22.3%	88.2%	78.5%	3.9

time factors (RTFs) were measured on a Linux workstation with a 500 MHz Pentium III processor. The baseline word recognizer does not include any boundary information; silence periods are ignored in the LM, and filled pauses are treated like words. This setup has been shown to yield optimal performance on this data set when no boundary information is available.

The WER for the integrated approach *without* additional prosodic features is about 4% (relative) lower than that of the baseline system. Furthermore, the boundary information is produced with a precision and recall rate of about 75% for both. For a comparison, we evaluated the prosodic classifier that is integrated into the VERBMOBIL (VM) system (cf. Section 1) on the word chains (after removing the boundary labels) that were produced by the integrated approach. (These word chains contain less errors than those produced by the baseline recognizer. We wanted to exclude this source of errors for the VM prosodic classifier, however, to directly compare the M3 classification performance with that of the integrated word-and-boundary recognizer.) This module uses an MLP classifier based on a set of 276 prosodic features combined with an n-gram language model (Kompe, 1997). The results of this sequential approach are almost identical with that of the integrated approach, which, at this stage, does not make use of any prosodic features.

In the following line, the results for our MLP–HMM hybrid architecture are given, which is based on a set of 24 prosodic features calculated on the basis of the energy and F_0 contour. No further improvement in word accuracy is obtained compared to the integrated system without additional prosodic features, but a slight improvement in the overall M3 classification performance: recall is improved by 1.2%, whereas precision is degraded by only 0.4%. This improvement is not statistically significant, however.

To evaluate the approach in the case of an ideal MLP classifier for M3 boundaries, we used the goal function for the MLP training both during training and recognition (see Section 7). This result can be regarded as an upper limit for improvements that can be achieved by optimizing the prosodic classifier within the given architecture, and without modifying the word HMMs. The drastic improvement in boundary classification rate is not surprising, because information about the position of phrase boundaries in the test sample is incorporated in this approach. Furthermore, a further relative reduction of WER by about 3% is obtained. This result indicates that additional knowledge about the position of phrase boundaries does improve word recognition, but it does not do so dramatically.

The results show that syntactic–prosodic phrase boundaries can be recognized within a regular HMM-based speech recognizer with a precision similar to that of a complex prosodic classifier operating on the word recognizer output. Furthermore, an improvement in word accuracy is observed compared to the pure word recognizer, which indicates that the integrated language model is more suitable for spontaneous speech than traditional word-based models.

The further improvement that is achieved by the integration of prosodic features is too small to justify the computational overhead involved. We are convinced, however, that a further improvement of the frame-based prosodic feature set and the MLP training is possible. For example, no durational information has been integrated to date. The results achieved by using the 'ideal' MLP output show that there is still room for improvements.

The integrated approach requires considerably less computational and memory resources, especially if no additional prosodic features are used. Compared to the sequential approach, a speed-up of 11% is achieved. The memory requirements of the integrated word-and-boundary recognizer are almost identical to those of the traditional word recognizer, whereas the prosody module requires an additional language model as well as additional HMM and MLP parameters. Finally, the integrated approach for word-and-boundary recognition can be employed with little additional program code if an HMM-based continuous speech recognizer is available, whereas the implementation of the prosody module required a considerable amount of time and effort.

A number of important questions have not been addressed in this paper and require further research. For example, we did not yet investigate how the use of different filled pause versions for boundaries and non-boundaries affects the recognition result. Also, we did not yet evaluate the optional introduction of phrase boundary symbols at word boundaries *without* consuming a time frame (see Section 4). It is also questionable whether the phrase boundary labels employed for our experiments are optimal for our purpose, as these have been originally designed for the sole purpose of improving the linguistic analysis process. It might be possible that an automatic procedure for determining suitable boundaries in the training data might yield a further improvement in word accuracy.

9. Summary and conclusion

In this paper, we presented the first integrated approach for the recognition of words and prosodic phrase boundaries. Whereas speech recognizers typically use language models that regard spoken utterances as unstructured sequences of words, our approach uses a more sophisticated model that regards utterances as sequences of words and phrase boundaries. This approach has two main advantages compared to the traditional, word-based model. First, additional syntactic–prosodic information is determined by the speech recognizer which facilitates parsing and resolves syntactic and semantic ambiguities. Although this additional information is determined without any computational overhead, the quality of the prosodic phrase boundary classification is comparable to that of the separate prosodic classification module used in the VERBMOBIL prototype. Second, the integrated model yields a slightly improved word accuracy compared to the traditional word-based approach (when evaluated after removing the boundary labels from the result of the integrated recognizer).

Furthermore, we described how prosodic information can be incorporated into the approach by employing an MLP–HMM hybrid recognizer and a frame-based suprasegmental feature set. The largest relative improvement in word recognition could be achieved without prosodic information, simply by including models for phrase boundaries in the vocabulary and in the statistical language model. Introducing an MLP classifier for phrase boundaries based on suprasegmental energy and F_0 features slightly enhances the boundary classification performance, but does not further improve the word accuracy. An experiment with the ideal MLP output indicates that knowledge about the position of phrase boundaries in the test data only slightly improves the word accuracy compared to the integrated word-and-boundary recognizer which does not incorporate this information. There is, however, a large potential of further increasing the boundary classification performance by enhancing the prosodic feature set.

We conclude that integrated recognition of words and prosodic phrase boundaries is superior to traditional word-based speech recognizers for spontaneous speech recognition tasks. The approach does not only provide an elegant framework for dealing with pauses and non-verbal phenomena during speech recognition, but it also competes favourably with the conventional approach of first determining the spoken word sequence and then employing a separate classifier for prosodic phrase boundaries.

Acknowledgements

This work was funded by the DFG (German Research Foundation) under contract number 810

939-9 and by the German Federal Ministry of Education, Science, Research and Technology (BMBF) in the framework of the VERBMOBIL Project under the grant 01 IV 701 K5. The responsibility for the contents of this study lies with the authors.

References

Batliner, A., Kompe, R., Kießling, A., Mast, M., Niemann, H., Nöth, E., 1998. M = syntax + posody: A syntactic–prosodic labelling scheme for large spontaneous speech databases. Speech Communication 25 (4), 193–222.

Beccetti, C., Ricotti, L.P., 1999. Speech Recognition – Theory and C++ Implementation. Wiley, Chichester.

Bub, T., Schwinn, J., 1996. Verbmobil: The evolution of a complex large speech-to-speech translation system. In: Proc. Internat. Conf. on Spoken Language Processing, Philadelphia, PA, USA, Vol. 4, pp. 1026–1029.

Gallwitz, F., 2001. Integrated stochastic models for spontaneous speech recognition. Dissertation, Technische Fakultät der Universität Erlangen–Nürnberg, Erlangen, Germany.

Gallwitz, F., Batliner, A., Buckow, J., Huber, R., Niemann, H., Nöth, E., 1998. Integrated recognition of words and phrase boundaries. In: Proc. Internat. Conf. on Spoken Language Processing, Vol. 7, Sydney, Australia, pp. 2883–2886.

Heeman, P.A., 1999. Modeling speech repairs and intonational phrasing to improve speech recognition. In: Proc. IEEE Workshop on Automatic Speech Recognition and Understanding, Keystone, CO, USA.

Heeman, P.A., Allen, J.F., 1999. Modeling speaker's utterances in spoken dialog. Computational Linguistics 25 (4).

Iwano, K., Hirose, K., 1999. Prosodic word boundary detection using statistical modeling of Moraic fundamental frequency contours and its use for continuous speech recognition. In: Proc. Internat. Conf. on Acoustics, Speech and Signal Processing, Phoenix, AZ, USA, Vol. 1, pp. 133–136.

Jelinek, F., 1997. Statistical Methods for Speech Recognition. MIT Press, Cambridge, MA.

Kießling, A., 1997. Extraktion und Klassifikation prosodischer Merkmale in der automatischen Sprachverarbeitung, Berichte aus der Informatik. Shaker Verlag, Aachen.

Kompe, R., 1997. Prosody in Speech Understanding Systems, Lecture Notes for Artificial Intelligence. Springer, Berlin.

Kompe, R., Batliner, A., Kießling, A., Kilian, U., Niemann, H., Nöth, E., Regel-Brietzmann, P., 1994. Automatic classification of prosodically marked phrase boundaries in German. In: Proc. Internat. Conf. on Acoustics, Speech and Signal Processing, Adelaide, Australia, Vol. 2, pp. 173–176.

Kompe, R., Kießling, A., Niemann, H., Nöth, E., Schukat-Talamazzini, E., Zottmann, A., Batliner, A., 1995. Prosodic scoring of word hypotheses graphs. In: Proc. European Conf. on Speech Communication and Technology, Madrid, Vol. 2, pp. 1333–1336.

Lea, W., 1980a. Prosodic aids to speech recognition. In: Lea, W. (Ed.), Trends in Speech Recognition. Prentice-Hall, Englewood Cliffs, NJ, pp. 166–205.

Lea, W., 1980b. Speech recognition: Past, present, and future. In: Lea, W. (Ed.), Trends in Speech Recognition. Prentice-Hall, Englewood Cliffs, NJ, pp. 39–98.

O'Shaughnessy, D., 1992. Recognition of hesitations in spontaneous speech. In: Proc. Internat. Conf. on Acoustics, Speech and Signal Processing, San Francisco, CA, USA, Vol. 1, pp. 521–524.

Ostendorf, M., Veilleux, N., 1994. A hierarchical stochastic model for automatic prediction of prosodic boundary location. Computational Linguistics 20 (1), 27–53.

Price, P., Ostendorf, M., Shattuck-Hufnagel, S., Fong, C., 1991. The use of prosody in syntactic disambiguation. Journal of the Acoustic Society of America 90, 2956–2970.

Schukat-Talamazzini, E.G., 1995. Automatische Spracherkennung – Grundlagen, statistische Modelle und effiziente Algorithmen, Künstliche Intelligenz. Vieweg, Braunschweig.

Shriberg, E., Stolcke, A., 1996. Word predictability after hesitations: A corpus based study. In: Proc. Internat. Conf. on Spoken Language Processing, Philadelphia, PA, USA, Vol. 3, pp. 1868–1871.

Stolcke, A., Shriberg, E., 1996. Statistical language modeling for speech disfluencies. In: Proc. Internat. Conf. on Acoustics, Speech and Signal Processing, Atlanta, GA, USA, Vol. 1, pp. 405–408.

Stolcke, A., Shriberg, E., Bates, R., Ostendorf, M., Hakkani, D., Plauche, M., Tür, G., Lu, Y., 1998. Automatic detection of sentence boundaries and disfluencies based on recognized words. In: Proc. Internat. Conf. on Spoken Language Processing, Sydney, Australia.

Stolcke, A., Shriberg, E., Hakkani-Tür, D., Tür, G., 1999. Automatic detection of sentence boundaries and disfluencies based on recognized words. In: Proc. European Conf. on Speech Communication and Technology, Budapest, Hungary.

Vaissière, J., 1988. The use of prosodic parameters in automatic speech recognition. In: Niemann, H., Lang, M., Sagerer, G. (Eds.), Recent Advances in Speech Understanding and Dialog Systems, NATO ASI Series, Springer, Berlin, Vol. 46, pp. 71–99.

Veilleux, N., Ostendorf, M., 1993. Prosody/Parse scoring and its application in ATIS. In: Human Language Technology. Proc. of the ARPA Workshop, Plainsboro, pp. 335–340.

Wahlster, W., 1993. Verbmobil – Translation of face–to–face dialogs. In: Proc. European Conf. on Speech Communication and Technology, Berlin, Germany, Volume Opening and Plenary Sessions, pp. 29–38.

Warnke, V., Gallwitz, F., Batliner, A., Buckow, J., Huber, R., Nöth, E., Höthker, A., 1999. Integrating multiple knowledge sources for word hypotheses graph interpretation. In: Proc. European Conf. on Speech Communication and Technology, Budapest, Hungary, Vol. 1, pp. 235–238.

Wightman, C., Shattuck-Hufnagel, S., Ostendorf, M., Price, P., 1992. Segmental durations in the vicinity of prosodic boundaries. Journal of the Acoustic Society of America 91, 1707–1717.