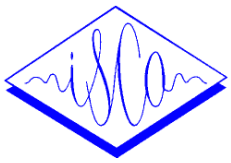

3rd International Conference on
Spoken Language Processing (ICSLP 94)
Yokohama, Japan
September 18-22, 1994

A CONTINUOUS SPEECH RECOGNITION SYSTEM INTEGRATING ADDITIONAL ACOUSTIC KNOWLEDGE SOURCES IN A DATA-DRIVEN BEAM SEARCH ALGORITHM

B. Plannerer, T. Einsele, M. Beham and G. Ruske

Lehrstuhl für Datenverarbeitung
TU München, Arcisstraße 21, 80290 München

ABSTRACT
The paper presents a continuous speech recognition system which integrates an additional acoustic knowledge source into the data-driven beam search algorithm. Details of the object oriented implementation of the beam search algorithm will be given. Integration of additional knowledge sources is treated within the flexible framework of Dempster-Shafer theory. As a first example, a rule-based plosive detector is added to the baseline system.

1. INTRODUCTION

Integrated beam search algorithms are well known to be efficient algorithms for finding the best sequence of words from a spoken sentence. During the search process, the best sequence of states corresponding to the sequence of feature vectors of the spoken input is determined. The states are usually identical with the states of the hidden Markov models of the acoustic knowledge source but could also be any pseudo-stationary segments of speech. For finding the best sequence of states, a scoring function is needed which provides a score $S(\vartheta_i^j, \vec{x}_t)$ for every state ϑ_i^j with state index i for every subword model j of the search space given a feature vector $\vec{x}_t$ for a given frame t. When using hidden Markov models, the emission probability density functions (or approximations of them) are used as scoring functions, but in hybrid systems e.g. the outputs of a neural net will provide the scores for every 'state'. Thus, the search algorithm operates on a fixed frame grid of about 10ms and uses an identical feature vector for every subword unit as input for the primary acoustic knowledge source. Some events in speech, however, could be detected more easily in a smaller frame grid with higher time-resolution (eg. 2ms). An example for these events are plosives, which usually are not well represented in a frame interval of 10ms. Therefore, one could build an event detector operating on such different feature vectors, which is specialized on detection of this event and may function as secondary acoustic knowledge source. One straight-forward method of integrating such detectors into the search process would require the following steps:

Synchronization of events:
Map the time frame at which the event is detected to the nearest 10ms frame of the global time raster.
Extension of the feature vector:
Add the event/no-event output of the event detector to the standard feature vector of the hidden Markov models.
Training:
Re-train the complete primary acoustic knowledge source using the extended feature vector.

In this paper we propose a different approach for integrating additional knowledge sources which will not require the complete retraining of the primary acoustic knowledge source and which will allow for easy adding / changing of the additional knowledge sources in experimental evaluations. In this approach, a rescoring of the state hypotheses of the primary knowledge source is performed with respect to the detector's output. This rescoring procedure may be carried out every frame or only at time instances when an event has been detected, depending on the quality of the secondary knowledge source. Thus, the primary knowledge source remains unchanged, and no retraining of the complete system is required. Furthermore, even a different training set could be used for training the new knowledge source, which might be useful in some cases when the training procedure for the additional knowledge source requires expert-knowledge and training cannot be performed automatically. As an example of such an additional knowledge source, we present a rule-based detector for plosives.
In section 2, we will present the baseline system, which was developed for the German 'ASL/VERBMOBIL' project. We will present the system architecture and some implementation details which allow an easy modification of the baseline system. In section 3 of this paper, the motivation and the theoretical framework for rescoring the state hypotheses based on the flexible Dempster-Shafer theory is presented. Section 4 will give some details of the rule-based plosive detector. In section 5 preliminary experimental results are presented.

2. ARCHITECTURE OF THE BASELINE-SYSTEM

Our system uses a data-driven beam search algorithm for finding the best sequence of states corresponding to a given utterance. The algorithm is based on a tree-structured pronunciation lexicon as proposed in [1]. In addition, a bigram language model is used. For further reduction of state hypotheses, a phoneme look-ahead is performed. The system architecture is depicted in Fig. 1. As mentioned, three knowledge sources are consulted by

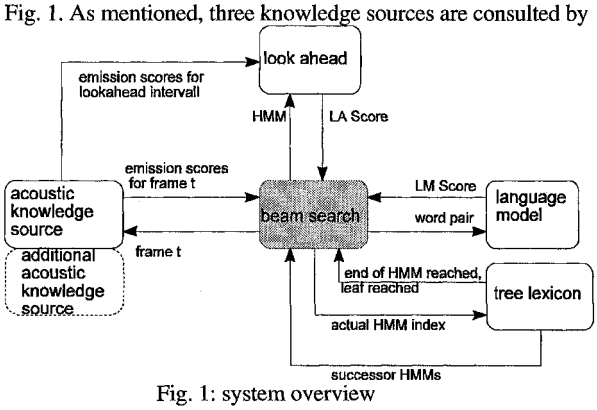

Fig. 1: system overview

the beam-search algorithm: the acoustic modeling, the tree-structured pronunciation lexicon and the bigram language model. Every knowledge source has been implemented as an independent object in C++. Communication between the search algorithm and the knowledge sources is done via functional interfaces to allow optimal encapsulation.

The search algorithm itself (including the phoneme look-ahead) controls all the search lists needed to span the search space driven by the actual data. We implemented the search lists as objects too, since this allows easy debugging and incorporation of statistical probes (eg. tracking the number of state hypotheses) into the search without changing the search algorithm itself. In the following, a brief description of the knowledge sources is given:

Tree lexicon:

For the ASL task, the vocabulary size is about 2000 words. Context-independent phoneme models are used as subword units for this task, but the algorithm allows any size and number of subword units. Every word may be described by an arbitrary number of pronunciations which will be represented in the tree structure. Using whole-word models for some frequent words is also supported. For the given vocabulary and using 43 context-independent phoneme models, a static reduction factor of 1.8 as compared to the linear lexicon could be achieved for the case of a single pronunciation for every word (note that the static reduction factor will increase for larger vocabularies, cf. [1])

Bigram language model:

For the ASL project, a bigram language model (statistical word pair grammar) was estimated by the Philips research laboratories, Aachen. Note that using a tree structured lexicon, recombination of paths is delayed by one word as compared to the linear lexicon and therefore tree copies are required in dependence of the predecessor word. As shown in [1], the number of tree copies is relatively small and due to efficient pruning, an overall speed-up can be achieved by using a tree lexicon.

As described above, every word may have several pronunciations. Therefore, when performing recombination on the nodes of the language model (i.e. on the word level), two cases have to be discerned:

- Two or more pronunciations correspond to one word: In this case, two or more leaves of the tree corresponding to the same word are reached and the path with the higher score is used for recombination on the node of the corresponding word.
- Two or more words are described by identical pronunciations: In this case, a leaf of the tree is reached that corresponds to two or more words. Thus, the path has to be expanded into every corresponding node of the language model and then normal recombination occurs.

In our implementation, every leave of the tree-lexicon is given an 'alias'-table containing the indices of all corresponding language nodes, whereby silences are treated as normal words.

Acoustic modeling:

The subword-units are based on context-independent phoneme models. Due to an object-oriented design of the models, semicontinuous or continuous hidden Markov models are both implemented, but only semicontinuous HMMs (SHMMs) were used for the experiments described in this paper.

The feature vector we defined consists of three groups of features: the first group is represented by the log-spectra computed by an auditorily based model, the second feature group is defined by the dynamic features (delta and delta-delta log spectra) and the third group contains the signal energy, delta and delta-delta energy and zero crossing rate.

Each of the three feature groups is modeled by an independently estimated codebook having 256 entries each, together forming a product codebook. The phoneme models consist of 3 or 4 states depending on the average length of the subword units, organized in the typical left-right structure allowing one skip transition. No transition probabilities were estimated; instead a fixed score for the transitions is added during the search, which did not affect recognition performance in our tests. The SHMMs were trained by the Viterbi training algorithm which we presented in [2], but no codebook reestimation was performed for the experiments given in this paper. Instead, the codebooks were simply estimated by a modified version of the K-means clustering algorithm.

Search algorithm:

Implementation of a data-driven beam search algorithm requires a complex structure of search lists organizing the search space in order to ensure strictly linear dependence of computation time and the number of search hypotheses. Our implementation of the search algorithm is strictly hierarchical on the level of the states, the subword units and the trees. On every level, path expansion and recombination occurs seperately. First, the expansion and recombination of paths is performed on the level of states. Expansion and recombination of paths is performed by forward-expansion of all the paths for the activated states into a recombination array, consulting the acoustic knowledge source. This results in a new list of active states for the next frame (after pruning). From the list of states, the communication list of ending HMMs is generated.

On the level of HMMs, this list will be used for starting the successor HMMs and for generating the list of reached leaves by consulting the tree lexicon. On the level of trees, the list of reached leaves is expanded according to the aliases of every leave and recombination on the nodes of the language model is carried out consulting the language model. Then, the tree startup follows and the expansion of paths will be performed level by level in descending order. By this organization, communication lists provide the neccessary information exchange between the individual levels. As mentioned above, every access to the knowledge sources (eg. requesting the score for a given state) is carried out via functional interfaces. Therefore, easy replacement or expansion of the knowledge sources is possible without modifying the search algorithm. On the level of trees, we introduced an additional pruning step after recombination on the language nodes to further reduce the number of tree startups. This is achieved by pruning those language nodes that are likely to fall below the pruning threshold after expansion of the corresponding HMMs. Of course, this pruning step may introduce additional search errors and is therefore not consistent. In our experiments, however, it showed no significant decrease of performance. The number of HMM-startups is further reduced by a phoneme look-ahead procedure preventing those phoneme models to be started that are likely to fall below the pruning threshold within the look-ahead interval.

Performance evaluation on the ASL-task (7000 sentences uttered by 100 speakers for training, 500 sentences of a different domain, uttered by 10 different speakers at significantly higher speed for testing) with the system described above showed a word recognition rate of about 76.7%, which can be valued as good results for this difficult task.

3. ADDITIONAL KNOWLEDGE SOURCES

For integration of additional knowledge sources, several methods are possible. As mentioned above, a straight-forward approach would be the extension of the feature vector of the HMMs by adding the output of the additional knowledge source. This

requires a complete retraining of the system, and in some cases, the modeling approach of the HMMs might not fit well to the output of the detector. An event detector for example, might have a binary output which would badly fit to mixture densities.
The approach presented in this paper differs from the previous in rescoring the scores of the primary knowledge source, leaving the primary knowledge source unchanged.
In the following we present a very general framework for rescoring the state hypotheses which is based on the Dempster-Shafer (D-S-) theory [3,4]. D-S- theory allows easy treament of various kinds of knowledge sources and is also capable of dealing with rejections ("no decision") of single knowledge sources. The Bayesian approach is contained as a special case in D-S-theory. Then the special case of the event detector is presented.
At first, a short introduction into D-S-theory shall be given: Later, we will apply it to the special case of the event detector.
Dempster-Shafer theory can be seen as a useful framework for combining multiple outputs of different knowledge sources. Given a common "frame of discernment" Θ, each knowledge source may assign a degree of belief $m(A)$ to every subset $A \subset \Theta$. $m(A)$ is called a basic probability number since it expresses the belief committed to exactly the subset A. The function $m(.)$ is called a "basic probability assignment" (BPA). Note that there are $2^{|\Theta|}$ different subsets that can be assigned a portion of belief, compared to $|\Theta|$ elements as in the Bayesian approach. The following normalization condition must hold: $\sum_{A \subset \Theta} m(A) = 1$
The empty set is always assumed to have a zero assignment: $m(0) = 0$.
To obtain the total belief committed to $A \subset \Theta$, all assignments for all subsets $B \subset A$ will have to be added to form the belief function $Bel(A) = \sum_{B \subset A} m(B)$.
Therefore, $Bel(0) = 0$ and $Bel(\Theta) = 1$.

Those subsets $A \subset \Theta$ with $m(A) > 0$ are called "focal elements" of the belief function. A focal element that consists of exactly one single element of Θ is called a "singleton".
Now for two BPAs $m_1(A)$ and $m_2(B)$ of two different knowledge sources, the combined basic probability assignement is obtained as the orthogonal sum of the original BPAs (Dempster's rule of combination):

$$m(A) = m_1 \oplus m_2 = \frac{\sum_{A_i \cap B_j = A} m_1(A_i) \cdot m_2(A_j)}{1 - \sum_{A_i \cap B_j = 0} m_1(A_i) \cdot m_2(A_j)}$$

From this new BPA, a new belief function may be computed.
Multiple knowledge sources may be recursively combined by the following rule:
$m^* = m_1 \oplus m_2 \oplus \ldots \oplus m_K = (m_1 \oplus m_2 \oplus \ldots \oplus m_{K-1}) \oplus m_K$
An interesting feature of D-S-theory is the easy treatment of rejections ("no decision") of single knowledge sources by assigning a rejection support S_{reject} to Θ: $m(\Theta) = S_{reject}$.
Thus, if the rejection support of a knowledge source is equal to 1, the combined BPA will not be influenced by this knowledge source. In Bayesian approach, the rejection must be expressed by assigning identical values $|\Theta|^{-1}$ to all elements of Θ.
In our case, we would like to define our frame of discernment to be the total set of states of all models of the acoustic knowledge source, e.g., all states of all HMMs in our baseline system: $\Theta = \{\vartheta_i^j\}$; $i = 1 \ldots N_j$; $j = 1 \ldots M$.
Then the basic probability assignment of the primary acoustic knowledge source would be the a-posteriori probabilities of the vector $\vec{x}_t$ being produced by the state ϑ_i^j of HMM with index j. This a-posteriori probability can be approximated by appropriate normalization of the scoring function (probability density function) $S(\vartheta_i^j, \vec{x}_t)$. Since the pseudo-a-posteriori probabilities assign an individual degree of belief to exactly every state, our primary knowledge source will only support singletons. Now for an additional knowledge source, various forms of basic probability assignments are possible. A broad phonetic classification, for example, would support subsets of the phoneme models but would not discriminate between the models that are members of the supported subsets. For a rejection, the secondary knowledge source would simply support the whole frame of discernment. Using multiple knowledge sources will result in a various number of supported subsets that can be combined step by step by Dempster's rule of combination. Note that even if the intermediate BPA's will support various subsets of Θ, the final combination with the HMM-classifier will support singletons only, since the primary knowledge source supports only singletons.
In the special case of event detection, the frame of discernment will be divided into two subsets: those state hypotheses which belong to the event to be detected, and those which do not correspond to the event. For example, a plosive detector would support the states of the phonemes /p/,/t/,/k/ as the first subset and would support all the other states of the other models as the second subset. That is, there is no further discrimintion between the individual plosives or even their corresponding states. In general, for event detection, the following sets can be defined:
let $\Theta = \{\vartheta_1^1, \vartheta_2^1, \ldots \vartheta_i^1, \ldots \vartheta_{N_1}^1, \ldots, \vartheta_1^j, \ldots, \vartheta_{N_j}^j, \ldots, \vartheta_1^M, \ldots, \vartheta_{N_M}^M\}$ be the frame of discernment , where $i = 1 \ldots N_j$ denotes the state index of model j and $j = 1 \ldots M$ denotes the model index. Further define $E = \{\vartheta_i^j\}$, $i = 1 \ldots N_j$, $j \in \Omega$ with Ω defining the set of those model indices corresponding to the detected event. Then, only three sets will be supported by the event detector: The set of event-related states E, the set of not related states $\neg E$ and the total set Θ.
For these three sets, the basic probability assignment of the event detector will be given as:

$$m(B) = \begin{cases} a, \text{ if } B = E \\ b, \text{ if } B = \neg E \\ 1-(a+b), \text{ if } B = \Theta \\ 0, \text{ else} \end{cases}$$

Using Dempster's rule of combination, the new belief function can be computed. Now for this case, a very simple rule of combination is found:

$$m^*(\vartheta_i^j) = \begin{cases} C \cdot m_1(\vartheta_i^j) \cdot (1-b) \text{ , if } \vartheta_i^j \in E \\ C \cdot m_1(\vartheta_i^j) \cdot (1-a) \text{ , if } \vartheta_i^j \in \neg E \end{cases}$$

with $m_1(A) = m_1(\vartheta_i^j)$ being the scoring function $S(\vartheta_i^j, \vec{x}_t)$ as described above. The normalizing factor C will represent the denominator term of Dempster's rule of combination and the normalizing factor necessary to obtain pseudo-a-posteriori probabilities from the scoring function $S(\vartheta_i^j, \vec{x}_t)$. Since during the search process, only maximum decisions between states will occur, this constant may simply be left out of consideration.

Thus, only a simple product term will have to be computed in this special case. However, for multiple knowledge sources, or if more complicated subsets are supported by the knowledge sources, the combination scheme may require a significant amount of computation.

4. PLOSIVE DETECTOR

An acoustic phonetic expert system generates the possibility of voiceless plosives using additional features which are extracted from the time domain.

Feature extraction

For a broad classification of speech sounds it is not necessary to evaluate the exact spectral shape. It has been shown that the root mean square (rms) energy in certain frequency subbands and the rate of rise (ror) of the log rms energy is sufficient to describe broad phonetic properties [5,6]. The basic inputs for our plosiv detector are the log rms energy of the outputs of three IIR (infinite impulse response) 6th order Butterworth bandpass filters with the following -3 dB bandwidth specifications: BP1 150 - 500Hz, BP2 1500 - 2500 Hz, BP3 2500 - 3500 Hz. These signals are directly computed in the time domain. A psychoacoustic motivated smoothing of the log energy is used to keep fast dynamic rises in the energy contour. After this smoothing the delta features - the time derivates of the log rms energy of the three bandpass filters - are calculated with the sampling rate of the speech signal and then reduced to the frame rate of the system (100 Hz) by detecting the maximum rate of rise in a 10 ms window.

Rule system

The additional features are evaluated in a acoustic-phonetic expert system using methods of fuzzy logic. All relations (big, small, rising, falling) are modeled with trapezoid membership functions. The $\wedge$ (and) and $\vee$ (or) operators are realized with min and max functions.

The following rules describe the three distinguishable segments of a voiceless stop consonant:

- **Begin of closure** ($B(t)$)**:** The log rms energy of all bandpass filters falls during the last 30 - 50 ms.
- **During the closure** ($C(t)$)**:** The log rms energy of BP1 and BP2 must be small and the spectral shape is flat.
- **End of Closure** ($E(t)$)**:** The ror of BP2 is high (burst detection), the total rms energy rises during the next 30 - 70 ms and the zero crossing rate is rising.

These three acoustic events must occur in a correct temporal order. The length of a plosive can be modeled with a trapezoid membership function which is the possibility of the length of the closure:

- **Length of closure** ($L(\lambda)$)**:** The lower limit of the possibility of the length is between 20 and 40 ms, the upper limit is between 100 and 200ms.

The final output of the plosive detector - i.e. the possibility of a plosive in frame t -

$$Plosive(t) = \bigcup_{\lambda} \bigcup_{i} PL_{\lambda}(t-i)$$

$$PL_{\lambda}(t) = B(t) \wedge E(t+\lambda) \wedge L(\lambda) \wedge \bigcap_{i=0}^{\lambda} C(t+i)$$

must be calculated for every time instant t in the speech signal and for every possible length λ (up to a mximal length λmax) of the plosive segment.

5. FIRST EXPERIMENTS

In our first experiments, the rule-based plosive detector as described above was used to rescore the state hypotheses of the baseline system. Since the plosive detector was originally designed for another system, its reliability in our application was relatively low. That is, the a-posteriori probability of being in a plosive state if an detection was given by the plosive detector, was only about 0.52. Therefore, using the difficult speech data of the ASL-Test, the overall perfomance could not be significantly enhanced, but as shown in Fig.2, even under that hard conditions, a reduction of the search space was achieved. It can be seen from Fig. 2 that a significant reduction of state hypotheses as compared to the baseline system occurs at those time instants where a plosive has been correctly detected (in this experiment, rescoring of state hypotheses was performed only at frames corresponding to a detection event).

Thus, even with relatively poor performance of the additional knowledege source, its information may be utilized during the search process .

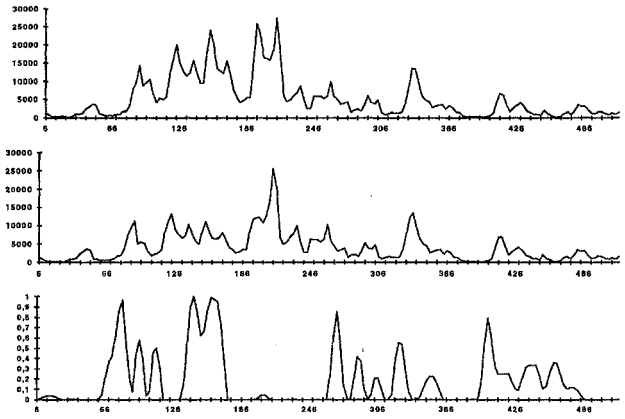

Fig. 2: a) number of state hypotheses of the baseline system,
b) state hypotheses after combination of knowledge sources,
c) output of the plosive detector

*This work has been carried out within the **ASL/VERBMOBIL** project of the German **BMFT**.*

6. REFERENCES

[1] H. Ney et al., "Improvements in beam search for 10000-word continuous speech recognition", Proc. of. ICASSP-92, March 1992, pp. I-9 -12.

[2] B. Plannerer and G. Ruske, "Recognition of demisyllable based units using semicontinuous hidden Markov Models", Proc. of. ICASSP-92, March 1992, pp. I-581 - 584.

[3] G. Shafer, "A mathematical theory of evidence", Princeton University Press, Princeton/London, 1976.

[4] Lei Xu et al, "Methods of Combining multiple classifiers and their applications to handwriting recognition", IEEE Trans. SMC, Vol.22, No. 3, May/June 1992, pp.418-435.

[5] L.F. Weigelt, S.J. Sadoff, J.D. Miller: Plosive/fricative distinction: The voiceless case. JASA. 87(6), June 1990, pp. 2729-2737.

[6] Y. Bengino, R. De Mori, et al.: Phonetically Motivated Acoustic Parameters for Continuous Speech Recognition using Artificial Neural Networks. EuroSpeech 91, Sept.1991, pp. 551-554.