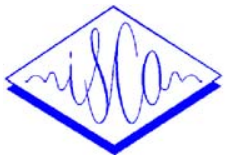

4th International Conference on Spoken Language Processing (ICSLP 96)
Philadelphia, PA, USA
October 3-6, 1996

Speech Recognition for an Information Kiosk*

J.L. Gauvain, J.J. Gangolf, L. Lamel
Spoken Language Processing Group
LIMSI-CNRS
91403 Orsay, FRANCE
{gauvain,gangolf,lamel}@limsi.fr

ABSTRACT

In the context of the ESPRIT MASK project we face the problem of adapting a "state-of-the-art" laboratory speech recognizer for use in the real world with naive users. The speech recognizer is a software-only system that runs in real-time on a standard Risc processor. All aspects of the speech recognizer have been reconsidered from signal capture to adaptive acoustic models and language models. The resulting system includes such features as microphone selection, response cancellation, noise compensation, query rejection capability and decoding strategies for real-time recognition.

1. INTRODUCTION

In this paper we address issues that must be faced in adapting a "state-of-the-art" speech recognizer developed in a laboratory for real-world use. All aspects of the speech recognizer must be reconsidered from signal capture to adaptive acoustic and language models. We have confronted these issues in the context of the ESPRIT MASK (Multimodal-Multimedia Automated Service Kiosk) project, aimed at providing access to rail travel information[6]. The speech recognition requirements for the MASK information kiosk are: speaker-independence; real-time spontaneous, continuous speech recognition; a recognition vocabulary of 1500 words, including almost 600 station/city names; and robustness to noise as the expected backgound noise level for the MASK kiosk located in a Parisan train station is on the order of 63dBA SPL.

In order to better simulate the acoustic conditions of the final kiosk, at LIMSI we built a data collection kiosk according to the physical specifications supplied by ergonomics experts[4]. This data collection kiosk, shown in Figure 1, is being used to carry out laboratory experiments prior to the availability of the final MASK prototype. The touch screen (1) is located so as to accomodate a wide variety of user heights, as per the recommendation of ergonomic experts. On the top (2), left (3) and right (4) of the screen are 3 microphones placed to allow for different heights and positions of users. This data collection kiosk allows us to carry out measurements and to record data under more realistic conditions, by placing the users in conditions closer to that of real use.

The spoken language system runs on a standard RISC workstation (Silicon Graphics Indy) with a standard UNIX operating system. This choice allows easily modification of the system, by simply transfering new executable versions to be incorporated in the kiosk. This is the same design choice as was taken for the MASK prototype, where communication with other system components is via an Ethernet connection. The MASK kiosk will have multimedia output, as well as a numeric keyboard and card slot for simulation of credit card payment, and a printer for information and tickets.

Figure 1: The LIMSI MASK data collection kiosk, (1) touch screen, (2), (3) and (4) are microphones, and (5) loudspeaker.

The MASK spoken language system[10] consists of a speaker-independent continuous **speech recognizer** who outputs the most probable word sequence or a word graph, which is passed to a **natural language** (NL) component. The NL component is concerned with understanding the meaning of the spoken query and includes the **semantic analysis**[3] and **dialog management**. Natural language responses are generated from the semantic frame, the dialog history and retrieved DBMS information. The text is typed on the screen and may be accompanied by other visual information (tabular form, ticket, etc) and/or vocal feedback using concatenated speech from stored dictionary units.

The continuous speech recognizer is a software-only system (written in ANSI C) that runs in real-time on a standard Risc processor. The system is independent of the speaker, so that no speaker-specific

*This work was partially financed by the ESPRIT project 9075 - MASK.

10.21437/ICSLP.1996-220

enrollment data is needed for a new user. Speaker-independence is obtained by training the acoustic models on speech data from a large number of representative speakers, covering a wide variety of accents and voice qualities. The recognizer uses continuous density HMM with Gaussian mixture for acoustic modeling and a *n-gram* backoff language models[14]. By using statistical language models, the user is not constrained to speak in complete sentences nor to to adhere to a preconstrained syntax. The system is evidently also able to recognize short phrases or isolated words. The recognition vocabulary for the MASK task currently contains about 1500 words, including 580 station names.

We are currently using the LIMSI kiosk to collect data with speech input only. A touch-to-talk mode is used, where subjects are asked to keep their hand on the screen while they are talking. WOz studies[17] found that subjects did not object to touch-to-talk, which substantially simplifies the work of the speech detection. This is important as if the system is continuously listening it needs to differentiate queries directed at the system from those directed at a traveling companion. Later rounds of data collection will allow both vocal and tactile inputs, in which case touch inputs will be mapped into the same semantic frame representation as used for spoken inputs.[1]

2. SIGNAL CAPTURE

Acoustic signal capture is an important design consideration. In our laboratory systems, we typically collect data using two microphone channels, a close-talking, noise cancelling microphone and a tabletop PCC microphone. However, for an information kiosk the microphones must be fixed, and must take into account various customer heights and positions when using the kiosk. We have chosen to position 3 PCC microphones around the screen cavity perpendicularly to the screen (see fig. 1). Based on the SNR of each channel, the output of one of the three microphones is selected. The speech signal is bandlimited to 8kHz and sampled at 16kHz. Beam forming was considered but found to not be efficient for the kiosk configuration, since the distance between the speaker and the closest microphone is less than the distance between microphones. A fourth channel is used to capture the signal played over the loudspeaker, coming from the message synthesizer or from video soundtracks, in order to compensate for the acoustic feedback on the microphones.

Measurements were carried out in a Parisian train station to estimate the expected mid working day background noise. Over a 2 hour period the average noise with a sound meter was found to be 63dBA SPL. Using the 3 microphone configuration without additional protection such as an acoustic hood,[2] we obtain an SNR of 18dB[3] if the customer is looking at the screen and is close enough to touch it.

In order to allow the user to start talking while the system is playing a message, a speech response cancellation module[19] was implemented. The kiosk impulse response at each microphone is modeled with an adaptive FIR filter, controlled by the Normalised LMS algorithm. Speech recognition experiments were carried using the response canceller in order to estimate the interest of such costly processing. We found that use of the algorithm improved the word accuracy when the user started speaking during the response. However, an even better result was obtained by cancelling the response signal only until the user's speech was detected, and stopping the response signal as soon as possible. The reason for this is that the kiosk's impulse response is dependent on the user's position, and that most users move while speaking, making filter estimation very inaccurate. We therefore only kept the speech detecter which given the response signal is able to accurately detect when the user has started to speak. Even though signal capture is continuously performed, a touch-to-talk mechanism is used to get a rough estimate of the query endpoints, as well as to avoid processing queries not directed to the system.

3. ACOUSTIC MODELS

We experimented with two front ends (MFCC and PLP based) and various configurations to find the best suited to our problem. Without use of a noise compensation technique[11], we found the PLP cepstrum[13] to be somewhat more robust to background noise than classical MFCC, i.e. the word error was reduced by 10%. Presently our noise compensation scheme is based on MFCC features and had not yet been ported to PLP cepstrum, so for noisy conditions we still use a classical MFCC analysis for the training data recorded in a quiet environment. More experiments are needed to compare both analyses with noise compensation. The PLP signal processing consists of a Mel-scale spectrum (no preemphasis, 21 triangular filters) computed on the 8kHz band using a 30ms frame window and a 10ms frame rate. This is followed by a root-LPCC analysis[1] yielding 13 cepstrum coefficients. The feature vector used for acoustic modeling is composed of the 13 cepstrum coefficients, 11 delta cepstrum coefficients and 6 delta-delta cepstrum coefficients. This reduced feature set gives the same recognition word accuracy as the full 39 feature set. Data is collected with the kiosk in an office environment, with an background noise level of about 46 dBA SPL due to computer equipment and ventilation.

Acoustic modeling is based on continuous density HMM with Gaussian mixture. Different acoustic model sets have been used for experimentation, comparing performance (speed and accuracy) using context-independent (CI) and context-dependent (CD) models sets, as well as gender-specific and speaker-independent models. When CD phone models are used, the contexts are automatically selected based on their frequencies in the training data. The contexts are independent of the word-position. When there is not enough training data to model a given triphone context, we backoff to right- and left-context phone models, and CI phone models.

The acoustic models were trained on 15k utterances from 300 speakers collected using interim versions of the spoken language system. The majority of the data was recorded at LIMSI (12k sentences from 194 speakers), with a smaller amount (3412 queries, 121 subjects) recorded by LIMSI at the Gare St. Lazare in Paris during the period of the MASK WOz experiments[16, 17].

[1] Even when both tactile and vocal inputs are available, we do not allow simultaneous speech and touch input within a single query. WOz studies[17] found that this is not an important limitation, as subjects tend to prefer a single mode (speech or touch), and even when they changed modes, they almost never did so within a single utterance.

[2] For the MASK kiosk different types of sound isolation are being tested - an enclosed cabin (like a telephone booth, a semi-closed cabin, and an isolating hood).

[3] The SNR definition used here, is the ratio of the averaged short term energies of the speech signal and the noise measured on 30ms windows after preemphasis with the following filter $1 - 0.95z^{-1}$.

4. LEXICON AND LANGUAGE MODELS

The Mask task recognition vocabulary currently contains about 1500 words, including 600 station names selected to cover the SNCF commercial needs. Except for the station names, the word list contains all words occuring at least twice in the training data. With this lexicon, the out-of-vocabulary (OOV) rate on the development test data is 0.6%. The lexicon is represented using a set of 35 phones. Frequent pronunciation variants are included in the lexicon, which also includes pseudo words such as "euh", "ah", "hum" and filler words such as "bon", "ben" as they are commonly observed in spontaneous speech.

Bigram and trigram language models have been estimated on the transcriptions of the training material, i.e about 15k utterances. A backoff mechanism[14] is used to smooth the estimates of the probabilities of rare n-grams by relying on a lower order n-gram when there is insufficient training data, and to provide a means of modeling unobserved n-grams. An advantage of the backoff mechanism is that language model size can be arbitrarily reduced by relying more on the backoff component, by increasing the minimum number of required n-gram observations needed to include the n-gram. This property is used in the first bigram decoding pass to reduce computational requirements. The trigram language model (LM) is optionally used in a second decoding pass. Word classes are used to provide better estimates for dates and times. The development set perplexities are 19 for the bigram and 16 for the trigram.

5. DECODER

An important aspect of real-time speech recognition is the design of a fast search algorithm that maintains high recognition accuracy. In the MASK system several techniques are combined, including a lexicon tree, multipass decoding, distributed LM weights, Gaussian shortlists and gender dependent (GD) acoustic models.

When using a bigram-backoff LM, decoding can be done with a static network, where the backoff component of the bigram is implemented with a tree organization of the lexicon[7]. In our system, the network is built in such a way that the word tails are shared between the lexicon tree and the linear representation of the words, so as to minimize the number of interword connections. With this implementation, the network size can be arbitrarily reduced by relying more on the backoff component. Bigram decoding with CI phone models is realized in real-time (RT), where real time is defined as taking 1s to process a 1s utterance. When a trigram LM is used, a second decoding pass is carried using a word graph generated with the bigram. We also use the result of the first decoding pass to guide the search of the second pass, and therefore can use a tighter pruning threshold. The second pass with more accurate acoustic and language models can be carried out in about 20% of CPU time of the first pass.

The language model weights are distributed over the phone graph so as to allow the use of a reduced pruning threshold, enabling both faster and more accurate search.

For small and medium vocabulary tasks, the state likelihood computation can represent a significant portion of the overall computation. One way to speed up this computation is to reduce the number of Gaussians needing to be considered to compute the likelihood for a state by preparing a Gaussian short list for each HMM state and each region of the quantified feature space[5]. Doing so, only a fraction of the Gaussians of each mixture is considered during decoding. This approach allows us to reduce the average number of examined Gaussians per mixture from 12 to 4 without any loss in accuracy.

One easy way to improve the accuracy of the recognizer is to use GD acoustic models. By building two separate networks and carrying out frame-synchronous decoding on the two networks in parallel, recognition can be improved without increasing the decoding time since after only a few frames the network corresponding to the speaker's gender is under consideration[15]. The small overhead of searching the 2 networks at the start of the sentence is largely compensated by more efficient pruning due to the use of more accurate models.

In passing from a laboratory system to an application an important need is the capability to reject out of domain queries. Our strategy is to estimate the a posteriori sentence probability for the recognizer hypothesis, i.e. $\Pr(w|x)$, by modeling the talker as a source of phones with phonotactic constraints provided by phone bigrams. We approximate $\Pr(w|x)$ by $\Pr(\phi_w|x) \simeq f(x|\phi_w)\Pr(\phi_w)/\max_\phi f(x|\phi)\Pr(\phi)$, where ϕ_w is the recognized phone transcription corresponding to the recognizer hypothesis w. $\Pr(\phi_w|x)$ is then compared to a fixed threshold to decide whether to accept or reject the query. This procedure requires only a small amount of additional computation.

6. NOISE COMPENSATION

Acoustic compensation is used in the recognizer to account for acoustic channel variability and background acoustic noise. We apply a data-driven model adaptation scheme as was used in the LIMSI Nov95 NAB system[11]. This adaptation is based on the following model of the observed signal y given the input signal x: $y = (x+n)*h$, where n is the additive noise and h the convolutional noise. In order to perform the speech analysis in real-time, sentence-based cepstral mean removal is replaced by removing the mean of the previously observed frames, where the cepstrum mean is updated at each frame with a first order filter (1 - $0.998z^{-1}$).

In order to better understand the effect of the acoustic environment on the performance of the speech recognizer, a series of experiments using noise recorded at the Saint-Lazare train station in Paris. The early version used for the experiments ran at 1.6xRT and had a word accuracy of 11.6% on a "clean" test data set having a signal to noise ratio of about 35dB. Various levels of SNR were simulated by adding noise at various amplitude levels to the test data (30dB, 24dB, 18dB and 12dB). The speech recognizer was evaluated with and without noise compensation.

SNR	*No compensation*	*With compensation*
30dB	11.3%	11.3%
24dB	12.6%	11.6%
18dB	19.2%	13.2%
12dB	42.5%	17.8%

Table 1: Word error as a function of the SNR.

Without compensation, the word error rate increases dramatically as the SNR is reduced. When the noise characteristics are known, the word error rate is seen to increase with the noise level, but the effect is less severe. Even in the worst condition (SNR of 12dB) the word error rate increases by only 50% compared to the 30dB SNR. From

these experiments we concluded that the speech recognizer performs at its maximum level with a 30dB SNR but that the performance may still be acceptable for a 12dB SNR. Based on these experiments a design objective of the MASK acoustic capture system (microphone setup, and acoustic isolation) was to obtain an SNR of 24dB. These experiments can only approximate the expected MASK conditions. The performance of the speech recognizer will depend not only on the SNR but also on the noise characteristics which can vary from station to station, and as a function of time. In addition, the speaking style of the users may also change as a function of the type and amount of noise. Finally, noisy conditions not only increase the error rate but also increase the recognition time. Field evaluations will be conducted when the first MASK prototype (ie. physical kiosk) is available, which will enable us to better estimate the effects of noise on the performance of the overall system.

7. RESULTS

Since word accuracy is very dependent on many factors not related to the acoustic data, such as the definition of a word, the out of vocabulary rate, and the language model, it is often easier to use a phone recognizer to compare acoustic model sets when trying to build the best set of models. However, when real-time recognition is a task constraint, performance may be more dependent upon other factors (such as the pruning level) than on the accuracy of the acoustic models, and the optimization procedure is a lot trickier. Our experience has been that improving the model accuracy not only improves recognition performance, but can also lead to better decoding due to more efficient pruning. However, if the decoding strategy remains the same, the trade off between accuracy and speed is dependent upon the total number of model parameters.

Our development strategy was to fix the constraint of real-time decoding and to find the best set of models given this constraint. We found that under this condition GD CI phone models outperformed GD CD phone models. The current best system configuration carries out decoding in 2 passes. The first pass decoding uses 2 sets of GD CI phone models (12 Gaussians/mixture) with a bigram LM (cutoff 1) to generate a word graph. The word graph, generated in real-time, contains on average 80 arcs per sentence and has a graph error of 4.7%. The second decoding pass is carried out with the selected set of 422 GD CD models (32 Gaussians/mixture) and a trigram-backoff LM (cutoffs 1 and 1). This decoding pass is carried out in 0.2xRT, and the resulting word error rate is 7.9% (the OOV rate of the test data is 0.6%). The use of CD models in the second pass reduces the word error by about 16%. This relatively small gain is surprising given that with the same CD models a phone error rate of 13.7% is obtained, compared to 24.3% obtained with a set of 35 CI models. This difference may be due to inadequecies in the language model.

About 20% of the errors are due to incorrect gender or number agreement, which are important for written French even though many of the words have the same pronunciation. If these errors (which are not important for understanding) are excluded the resulting word error rate is 6.3%.

8. CONCLUSIONS

This paper has described the work we have carrried out adapting our state-of-the-art laboratory speaker-independent, large vocabulary continuous speech recognizer for use in the MASK task. Signal capture is via multiple microphones, selecting the microphone with the highest SNR. In order to allow a natural interation with the machine, a response cancellation algorithm has been implemented so that the user can start talking during playback. Using a multipass decoding strategy, recognition is carried out in 1.2 x real-time with CD phone models and a trigram language model. To achieve real-time decoding distributed language model weights and Gaussian shortlists are used. Noise compensation and rejection capabilities are included to suit the needs of the MASK task.

9. ACKNOWLEDGEMENT

The authors acknowledge the contribution of Danial Solé to the development and testing of the speech response cancellation algorithm.

REFERENCES

[1] P. Alexandre, P. Lockwood, "Root cepstral analysis: A unified view. Application to speech processing in car noise environments," *Speech Communication*, **12**(3), 1993.

[2] L.R. Bahl et al., "A Fast Match for Continuous Speech Recognition Using Allophonic Models," *ICASSP-92*.

[3] S.K. Bennacef et al., " A Spoken Language System For Information Retrieval," *ICSLP'94*.

[4] F. Bernard et al., "Ergonomic Constraints on the Physical Design of the MASK kiosk," MASK *project deliverable 6.2*, June, 1994.

[5] E. Bocchieri, "Vector quantization for efficient computation of continuous density likelihoods," *ICASSP-93*.

[6] E. Chhor, I. Salter, "The MASK Project," *Human Comfort & Security Workshop*, Brussels, Oct. 1995.

[7] J.L. Gauvain, L. Lamel, "LIMSI Nov92 WSJ Evaluation," presented at the *ARPA Spoken Language Technology Workshop*, MIT, Cambridge, MA, Jan. 1993.

[8] J.L. Gauvain et al., "The LIMSI Continuous Speech Dictation System: Evaluation on the ARPA Wall Street Journal Task," *ICASSP-94*.

[9] J.L. Gauvain, L. Lamel, M. Adda-Decker, "Developments in Continuous Speech Dictation using the ARPA WSJ Task," *ICASSP-95*.

[10] J.L. Gauvain et al., "The Spoken Language Component of the MASK Kiosk," *Human Comfort & Security Workshop*, Brussels, Oct. 1995.

[11] J.L. Gauvain et al., "Developments in Continuous Speech Dictation using the 1995 ARPA NAB News Task," *ICASSP-96*.

[12] L. Gillick, R. Roth, "A Rapid Match Algorithm for Continuous Speech Recognition," *DARPA Sp&NL Workshop*, 1990.

[13] H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech," *JASA*, **87**(4), 1990.

[14] S.M. Katz, "Estimation of Probabilities from Sparse Data for the Language Model Component of a Speech Recognizer," *IEEE Trans. ASSP*, **35**(3), 1987.

[15] L. Lamel, J.L. Gauvain, "Continuous Speech Recognition at LIMSI," *DARPA Continuous Speech Recognition Workshop*, Sep. 1992.

[16] L. Lamel et al., "Development of Spoken Language Corpora for Travel Information," *Eurospeech'95*.

[17] A. Life et al., "Data Collection for the MASK Kiosk: WOz vs Prototype System," *ICSLP'96*.

[18] H. Murveit et al., "Large-Vocabulary Dictation using SRI's Decipher Speech Recognition System: Progressive Search Techniques," *ICASSP-93*.

[19] D. Solé. "Annulation d'écho pour un système de reconnaissance", LIMSI internal report, June 1996.