Speech Recognition over NetMeeting Connections

Florian Metze, John McDonough, and Hagen Soltau

Interactive Systems Laboratories
University of Karlsruhe, Germany
{metze|jmcd|soltau}@ira.uka.de

Abstract

In this paper we evaluate the performance of the ISL's German Verbmobil spontaneous speech recognizer on the Nespole! database. In this task, people talk to an agent in a tourist office to plan their holidays via a NetMeeting connection, also sharing screen contents (web-pages). Stereo recordings were made both before and after speech transmission over an IP connection using the G.711 codec, so that we are able to directly measure the loss in LVCSR performance due to NetMeeting's segmentation and compression. The aim of this work is to quantify this loss, which is a consequence of using protocols which were not designed for speech recognition purposes. We report on techniques employed to port our existing clean-speech recognizer to this new data quality, using about 1.5h of labeled adaptation data, but avoiding a complete retraining of the system.

1. Introduction

Microsoft's®NetMeeting™(currently in version 3.01) is a widely available and accessible front-end for voice transmissions over the Internet (VoIP). It is mainly used for human-to-human meetings, but as it also enables the participants of a video- or audio-conference to share other data, it lends itself to automated data collection or computer-assisted human-to-human interaction (translation, meeting summarization, etc). Human-to-computer interaction also seems possible, if one terminal is replaced by a software-only solution. For all these applications, speech will probably be the main modality.

Automatic speech recognition (ASR) on signals, which have been transmitted using current Internet standards faces several problems, mainly that the acoustic signal is usually transmitted in a compressed format and the fact that IP sends data in short packets instead of a continuous stream, often resulting in deletions.

If this speech is to be used both for re-synthesis (VoIP) and for automatic speech recognition (often called VRoIP), one is faced with the following dilemma:

1. For VoIP to be an effective means of communication between two humans, the transmission should be "instantaneous". Humans usually recover easily from a few dropouts (missing frames) during speech transmission; the efficiency of communication is not affected. However, people are used to getting answers immediately and find time delay in the transmission channel disturbing.
2. ASR engines usually do not run synchroneously with real time, but their performance is affected by bad segmentation and missing signals, even if only for a few milliseconds; on the other hand, it is easy to freeze the recognition engine, until new data arrives.

Several specialized solutions to VRoIP have already been investigated, but they rely either on on a tight integration of the speech recognizer with the transmission protocol [5], or a complex interface with proprietary systems [11]. The aim of this work is to present a speech recognition system for IP connections, that can be built by combining existing components, does not rely on detailed knowledge of the transmission protocol and is compatible with existing infrastructure, which is installed in most offices and many homes.

Especially, we do not want to retrain a recognizer from scratch, but investigate how an existing, state-of-the-art recognizer for spontaneous clean speech can be ported to the new task.

In the following section, we will outline the data set and our system setup, continue with a description of the speech recognizer used in this work, then discuss the different adaptation approaches pursued in this work and finally evaluate the performance of the overall system.

2. The Nespole! task

Nespole![1] is a joint EU[2]/ NSF[3] project aiming at providing multi-modal support for human-to-human interaction over IP networks.

During the initial data collection, an American, French, German or Italian client would call an agent at the Trento (Italy) tourist office and enquire about the holiday opportunities in the area. The agents were employees of the tourist office speaking the caller's native language. The callers were given material to allow them to choose from 5 scenarios for which material had been prepared, but were otherwise free to ask what they wanted. The aim of the Nespole! project is to develop showcases, which will provide multi-modal support for this process including gestures and machine translation.

For our experiment we used the "client" part of 43 dialogues between a client calling from Karlsruhe, Germany, and an agent. Figure 1 shows the typical setup used for recording this data. The stereo recordings were produced by combining local (high-quality) recordings with the remote recordings, which contain identical utterances, but after transmission over a real H323 channel.

Both parties in the conversation used standard, consumer-grade PCs running Microsoft Windows 98 or NT. The recordings in Karlsruhe were done in different offices, sometimes with people talking in the background, and using table-top microphones that come bundled with "Multimedia" computers.

[1] An acronym for "NEgotiation through SPOken Language in E-commerce", further information can be found at: http://nespole.itc.it

[2] European Union

[3] The United States' National Science Foundation

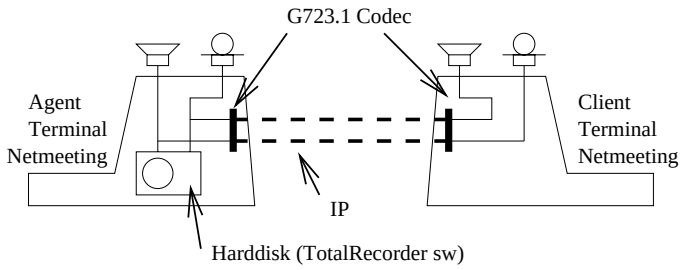

Figure 1: Typical setup for Nespole! recordings.

Recordings were done in 48kHz at the client's side and in 22.05kHz at the agent's side after transmission through the NetMeeting connection, which had been set to "CCIT u-law, 8kHz, 8bit" (G.711). For recognition, the signal files were re-sampled to 16kHz. To collect realistic data, recordings were scheduled at normal office hours, so that air-conditioning was on and traffic on the Internet was normal. Sound recording was done with the TotalRecorder[4] software, which has the capability to capture sound I/O directly from the PC's sound card and write it to a file. This file can then immediately be read in by the speech recognizer.

Set	Dialogues	Words	Turns	Duration
Adaptation	34	13,213	359	1:21h
Evaluation	9	2,901	1,879	0:25h

Table 1: Nespole! German adaptation and test set.

The recorded data was divided into several subsets, the characteristics of the two sets (adaptation and development test set) we used in this work are summarized in table 1. Adaptation and test set consist of 19 and 6 speakers respectively.

3. The ISL's Verbmobil-II recognizer

We conducted our experiments with the Interactive Systems Laboratories' recognizer for spontaneous German speech which we used in the Verbmobil-II system [9], described in detail in [8]. This recognizer uses our JANUS-III ASR toolkit and was trained on 62h of speech[5]. The software is written in C with an Tcl/Tk interface for scripting and currently runs on OSF/1 (Alpha), Linux (ix86), SunOS (Sparc), and Windows NT (ix86).

The Verbmobil-II recognizer employs 3,500 context-dependant models with 48 diagonal Gaussians each over a 32-dimensional LDA feature, computed from MFCCs and their deltas and delta-deltas. We also use semi-tied covariance matrices. During maximum likelihood training along pre-computed frame-state alignments, we used Cepstral Mean Subtraction (CMS), Vocal Tract Length Normalization (VTLN) and, for the last two iterations, speaker-dependant feature spaces. The computation of these feature spaces is based on a normalized likelihood function [3].

During the decoding stage, we employ these methods in an incremental, delayed way, starting from default values to adapt rapidly to new channels and speakers. To speed up the decoding process, we perform Gaussian selection using the BBI algorithm [2] and a context-independent phone Look-Ahead.

The language model is a tri-gram language model trained on 640k words from the Verbmobil corpus as well as transcriptions (approximately 17,000 words) from the Nespole! adaptation data. So far, we have not conducted experiments with interpolation and weighting of the different text corpora. The perplexity on the test-set was 98.5. The dictionary was expanded by 200 words to cover the new domain. The new pronunciations were generated by a rule-based approach. The new dictionary contains 11,800 different entries for a vocabulary of 11,200 words, resulting in an OOV-rate on the test-set of 1.6%.

Our system reached a word accuracy of 74.8%[6] in the final Verbmobil evaluation [4] and of 67.7% on the new task, the word accuracies for single dialogues range from 49.2% to 80.6%. We attribute this loss of performance to the new domain, which contains several foreign names, which several "clients" were not sure how to pronounce, as well as the different recording setup: the agent could send web-pages to the client's screen, and the clients usually reacted quite spontaneously when receiving new pages.

4. Adaptation experiments

Transmission of G.711 encoded speech over the Internet distorts the speech significantly and we tested several compensation techniques, namely Acoustic Mapping, Maximum A-Posteriori estimation and Maximum Likelihood Linear Regression (MLLR), to improve speech recognition. We also investigated the influence of lost packets occurring in real transmissions on speech recognition performance and discuss the role of segmentation.

Our standard training and decoding scheme includes dialogue-specific feature space adaptation (FSA) and VTLN. We first validated that these techniques work correctly also on the H323 data. The results are summarized in table 2.

Results without supervised adaptation	Clean	H323
Baseline	67.7%	30.8%
VTLN only	65.0%	27.2%
FSA only	65.5%	28.4%
Plain models	60.9%	24.5%

Table 2: Baseline performance of the Verbmobil-II system on the Nespole! data.

We also computed an FSA matrix on the adaptation data and loaded this matrix during decoding at the beginning of each dialogue instead of the standard matrix for either men or women, estimated on clean speech. The performance then improved from 30.8% to 31.5% word accuracy. This proves the good convergence of the ML adaptation scheme. In some cases, however, the noise contained in the H323 recordings prevented convergence of the ML criterion for the warping factor needed for VTLN, so that we had to follow a more conservative update strategy to obtain reliable estimates. We used this system to write initial labels for our supervised adaptation experiments

[4] `http://www.highcriteria.com/productfr.htm`

[5] The Verbmobil database consists of dialogues from the travel-planning and meeting arrangement domain.

[6] This system also used a context-independant phone Look-Ahead to reduce the number of score computations and increase decoding speed, which we did not use for the experiments described in this work.

using the Flexible Transcription Alignment technique described in [1].

We will now present the results achieved with the different adaptation methods.

4.1. MAP and Acoustic Mapping

Bandwidth is limited on most Internet connections, so that compression is generally used on these tasks. We tried to compensate for the effects of compression by transforming both the features and our acoustic models through MAP, MLLR and an Acoustic Mapping procedure similar to RATZ [6].

In our first experiments, we tried to mix the sufficient statistics from 62h of Verbmobil training data with the new data (1:20h). The results for different normalized weighting factors during MAP adaptation for the two sets of data are summarized in table 3.

Weighting factor	No MAP	0.4	0.6	0.8
Word Accuracy	30.8%	47.0%	49.0%	48.2%

Table 3: Results for MAP adaptation.

Using much simpler acoustic mapping as proposed in [6], we were able to reach a word accuracy of 42.6% by employing this technique in log-MEL space. We also observed a "normalization" effect, in that bad speakers improved more than good speakers. Acoustic Mapping works by calculating a soft partitioning of the two feature spaces using corresponding prototypes. In our case, we used codebooks containing one Gaussian for each context-independent speech state. We therefore know the transformation for each prototype; for an arbitrary vector, the transformation is calculated as the linear combination of the prototype's transformations where the weights are determined by the soft clustering using, in our case, 139 Gaussians with diagonal covariance matrices.

4.2. MLLR

MLLR proved to be the most effective single technique to adapt the recognizer. We use a hierarchical clustering scheme to avoid adapting codebooks on insufficient training data.

The best system used 48 32x32 adaptation matrices, which is determined by a min-count for updates of 8000, requiring that every adaptation matrix be trained on at least 80 seconds of data.

No. of Matrices	Baseline	89	48	35
WA	30.8 %	52.3%	53.1%	52.0%

Table 4: Word accuracy results for MLLR adaptation.

As we had also tried for MAP, we applied MLLR iteratively, by collecting new statistics with an adapted system and re-calculating the adaptation. However, these experiments did not lead to further improvement in recognition performance, nor did performing these experiments in a dialogue-dependant feature space. Again, we attribute these effects to the fact that our system has a rather high number of parameters, which are difficult to estimate with the amount of adaptation data we had, especially as a significant amount of the distortion we wish to compensate for is of non-stationary nature.

4.3. Missing frames

Standard H323 transmission does not recover lost packets, as it does not improve human-to-human speech transmission to resynthesize these packets at a later time. To compensate for this loss of information we implemented a simple detection algorithm for these missing frames and reconstructed the spectral features by using linear interpolation.

In our setup, the channel delivers no information on which packets are missing, so that a separate detection algorithm for bad frames is needed. After manual analysis of the audio data we suspected that the most detrimental effects are caused by regions that are characterized by a low zero-crossing rate, compared to the average zero-crossing rate for this signal file (ADC). There is no indication however, whether the artefact is solely an insertion of silence or indeed missing information. We therefore cut out silence frames and interpolate the partially affected frames at the border of the cut-out regions.

The detection algorithm works by comparing the zero-crossing rate of the ADC in a 10ms window to the average zero-crossing rate of the current utterance. If the ratio falls below an empirically determined threshold, the affected frames will be cut and interpolated. Using this approach, we improved the performance of our previously best system (WA 53.1%) to a word accuracy of 53.6%.

Other researchers have also used linear interpolation in conjunction with similarly encoded data, reporting successful reconstruction of missing frames [5]. However, in their work the position of the missing packets was known, so that no detection was necessary. Also, packets could be reconstructed before the audio file used for speech recognition was generated.

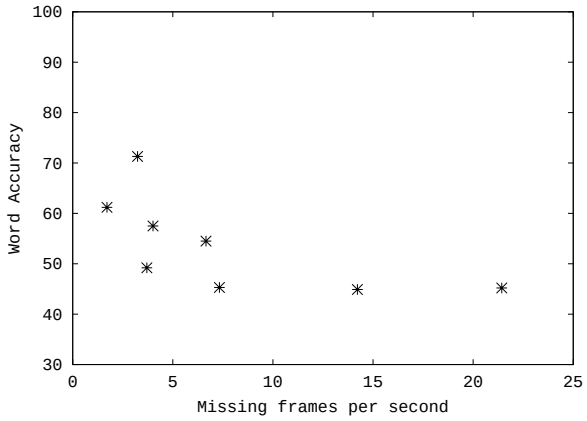

Figure 2: Influence of missing frames on recognition performance after reconstruction (1 second has 100 frames).

Figure 2 shows the effect of missing frames on the recognition performance of our recognizer. We see that the number of frames interpolated by our simple detection algorithm is related to recognition performance, although there is insufficient data to confirm whether a certain amount of missed frames can be tolerated without loss of performance. Dialogue-specific improvements in word-accuracy are more pronounced for "bad" dialogues, but we did not find a clear pattern in them and frame reconstruction also slightly increased the number of errors on two dialogues.

4.4. Reverse transformation

When testing the best adapted models on the original clean speech, the performance dropped to 59.8%, but by applying acoustic mapping as discussed in the previous section, a word accuracy of 64.4% could be reached, which compares favorably to the clean model's original performance of 67.7%. Acoustic Mapping therefore proves to be an effective technique despite the use of ML-FSA in our standard decoding setup. In this case, as MLLR does not adapt to non-stationary noise when we adapted our models to the H323 domain, the reverse mapping can be successfully done with simple acoustic mapping. It is therefore possible to use one recognizer to decode both clean and distorted speech without loading new acoustic models.

4.5. Segmentation

Contrary to our expectations, segmentation did not prove to be a major source of errors in the overall system. We expected one-word turns to be swallowed completely or at least be severely affected by NetMeeting's automatic segmentation. However, separate alignments of one-word-turns (like "mhm" or "ja") and first and last words in the reference and recognizer output did not show negative effects, when compared to the manual segmentation of the clean speech.

WA	Baseline	one word turns	first and last words
H323	53.1%	65.9%	55.1%
Clean	67.7%	80.2%	68.0%

Table 5: Influence of segmentation on ASR performance.

5. Discussion

We have shown how the performance of a clean-speech recognizer on an H323 task can be improved from 30.8% to 53.6% using several adaptation techniques. 50% of the loss in word accuracy compared to the recognizer's performance on clean speech has therefore been recovered.

Some of the turns were hard to understand even for humans; during one dialogue, 7 of 77 (9%) turns were repeated by the subjects, because the other party could not understand what had been said. The performance of our recognizer improved from 45.3% to 50.2% on this dialogue, if one does not count errors that had been irrecoverable even for humans.

Experiments conducted on a subset of the English Nespole! corpus, for which not only stereo recordings, but also separate human-produced transcriptions for these parallel recordings are available, have shown that there is a 3.0% difference in the transcriptions of these utterances. These 3% cover completely missing segments or genuine transcription errors, as no noises have been transcribed. The German data was transcribed to the same standards and, as we align the output of the speech recognition on the H323 data against the transcriptions of the clean speech, our numbers represent the overall performance of the system, including segmentation.

6. Conclusion

Our experiments have shown that existing speech recognizers can be made to work on real-world VRoIP problems, while being compatible with standard consumer hard- and software, given about one hour of labeled adaptation data. Several established adaptation techniques have been applied and compared and we showed how they can be employed to significantly improve the speech recognition performance on a 10k task, so that NetMeeting conferences could soon be integrated seamlessly into Meeting Recognition Systems, as described in [10] for example. A thorough analysis of distortion patterns of real speech after transmission over IP networks will certainly lead to further significant improvements. The prototype described in this work allows such a data collection to be implemented easily by using standard components on the user end.

7. Acknowledgments

This work was sponsored by the European Union under Grant No. IST1999-11562 as part of the Nespole! project. This project is part of the joint EU/ NSF MLIAM research initiative. The authors wish to thank all members of Interactive Systems Laboratories at the University of Karlsruhe and Carnegie Mellon University for useful discussions and support.

8. References

[1] Finke, M., and Waibel, A., "Flexible Transcription Alignment", in Proc. Automatic Speech Recognition and Understanding Workshop (ASRU); 1997.

[2] Fritsch, J., Rogina I.; "The bucket box intersection (BBI) algorithm for fast approximative evaluation of diagonal mixture Gaussians", in Proc. ICASSP 1996; Atlanta, USA; 1996.

[3] Gales, M. J. F., "Maximum likelihood linear transformations for HMM-based speech recognition", Technical report, Cambridge University, Cambridge, England; 1997.

[4] Malenke, M., Bäumler, M., Paulus, E.; "Speech recognition performance assessment", in Wahlster, W.; "Verbmobil: Foundations of Speech-to-Speech Translation", Springer; Heidelberg, Germany; 2000.

[5] Milner, B., and Semnani, S.; "Robust Speech Recognition over IP Networks", Proc. ICASSP 2000; Istanbul, Turkey; 2000.

[6] Moreno, P., Raj, B., Gouvea, E., and Stern, R.; "Multivariate-Gaussian-Based Cepstral Normalisation for Robust Speech Recognition", in Proc. ICASSP 1995; Detroit, USA; 1995.

[7] Raj, B., Seltzer, M., and Stern, R.; "Reconstruction of Damaged Spectrographic Features for Robust Speech Recognition", in Proc. ICSLP2000; Beijing, China; 2000.

[8] Soltau, H., Metze, F., Schaaf, T., and Waibel, A.; "The ISL evaluation system for Verbmobil-II", in Proc. ICASSP2001; Salt Lake City, USA; 2001.

[9] Waibel, A., Soltau, H., Schultz, T., Schaaf, T., and Metze, F.; "Multilingual speech recognition", in Wahlster, W.; "Verbmobil: Foundations of Speech-to-Speech Translation", Springer; Heidelberg, Germany; 2000.

[10] Waibel, A., Bett, M., Metze, F., Ries, K., Schaaf, T., Schultz, T., Soltau, H., Yu, H., and Zechner, K.; "Advances in Meeting Record Creation and Access", in Proc. ICASSP 2001; Salt Lake City, USA; 2001.

[11] Zhou, Q., Kosenko, S.; "Lucent Automatic Speech Recognition: A Speech Recognition Engine for Internet and Telephony Service Applications", in Proc. ICSLP2000; Beijing, China; 2000.