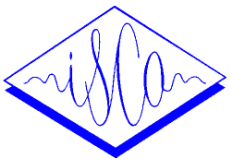

ISCA & IEEE Workshop on Spontaneous
Speech Processing and Recognition
Tokyo Institute of Technology, Tokyo, Japan
April 13-16, 2003

Development of a spontaneous speech recognition engine for an entertainment robot

H. Lucke, H. Honda, K. Minamino, A. Hiroe, H. Mori, H. Ogawa, Y. Asano, H. Kishi

Digital Creatures Laboratory, Sony Corporation
6-7-35 Kitashinagawa, Shinagawa-ku, Tokyo 141-0001, JAPAN

Abstract

Natural speech interaction is a difficult, yet important, capability for a social humanoidal robot. We address the problems of spontaneous speaking style in a real environment and report on our progress of developing a robust large vocabulary speech recognition engine for an anthropomorphic entertainment robot SDR-4X.

1. Introduction

Social and entertainment robotics is a relatively new field, that is rapidly gaining recognition. It is studied from a variety of perspectives ranging from developmental psychology to commercial applications. We have previously released a 4 legged entertainment robot AIBO (Fig. 1) which interacts with the user through motion, its vision and audio system, and tactile sensors. The first version did not contain speech recognition. Nevertheless, users, including those who knew that the device was not capable of recognizing speech, showed a strong tendency to communicate to the robot by voice.

We are currently working on a new prototype humanoid entertainment robot SDR-4X. Like its predecessor it has a repertoire of autonomous behaviors. It can be expected that users will feel a strong desire to communicate to the robot by voice. Therefore spontaneous open dialogue is an important goal to be realized on a humanoid platform.

It is generally agreed that spontaneous speech is far more difficult to recognize than read speech. Experiments of recognition rates of the same material when uttered spontaneously or read by the same speakers show a considerable difference [4]. The reason lies in the significant segmental and supra-segmental variations that can be observed in spontaneous speech [8]. What makes the problem so difficult is (1) a lack of understanding on what these variations are and how to model them and (2) a lack of accurately labeled training data as it is far easier to solicit read speech for recording purposes than it is to solicit spontaneous speech. In this paper we address the problem of spontaneous speech towards a robot in a real environment by focusing on the acoustic and linguistic training data and on environmental effects.

2. Spontaneous speech in a real environment

As strong as the need for natural speech dialogue on a social robot platform is, as large are the problems to realize it: Speech recognition is made difficult by environmental noises, actuator noses, room reverberation due to far-field microphone and a spontaneous, sometimes emphatic speaking style.

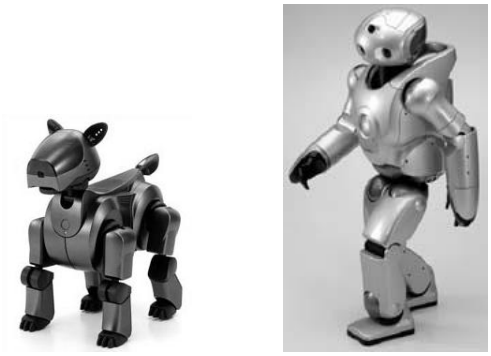

Figure 1: Autonomous entertainment robots: Commercial 4-legged robot AIBO (l); humanoid prototype SDR-4X (r).

2.1. Far field speech input and low signal to noise ratio

In the SDR-4X the main microphones are located on either side of the head. The speaker might well be several meters away from the microphone. This reduces the signal-to-noise-ratio significantly and also means room reverberations play a much bigger role. In addition background noises are a significant problem. Not only is an entertainment robot likely to be used in a noisy environment (e.g. background music or TV) but also the robot itself generates various noises, the noise source of which is usually at a much closer distance than the speaker. We measured SNRs ranging from 23.6 dB when the robot is standing idle on a carpet floor 50cm from the speaker, to -6.3 dB when it is shaking its head at a speed of 1 Hz at 3m distance.

2.2. Unconstraint spontaneous dialogue

Speech dialogue interfaces for a variety of robots have been proposed. However, most approaches focus on a task-oriented dialogue which is usually limited to a certain domain. Also a cooperative speaking style can be expected.

In contrast, we are interested in 'entertainment oriented dialogue' similar to chatting between friends. This is characterized by a casual spontaneous speaking style on a potentially unlimited domain.

The problem of "understanding" spontaneous speech after it has been recognized, in itself is a difficult problem. We hope that the SDR-4X will provide a platform for studying the speech understanding issues involved. In the short term, we believe, that through relatively simple pattern matching techniques, it is possible to generate responses that appear to make sense even in the absence of 'true' speech understanding. This is similar to the classic Eliza program which conveyed a certain feeling of text understanding[1].

[1] Eliza used text input instead of speech

	training data corpus	type	# speakers male	fem	total time [h]
a	IPA	read	183	187	54.84
b	sn601	read	32	32	8.45
	dra01	read	56	64	58.04
	VerbMobil	spont.	185	259	38.37
	TV	spont.	221	120	16.32
	total		494	475	121.20

Table 1: Statics of speech material used for training the baseline HMM (IPA9904) (a) and spontaneous speech HMM (Sony0204) (b).

3. Implementation on SDR-4X

With current technology it is not yet possible to obtain high recognition performance under the demanding constraints described above. Therefore we adopted an adaptive approach in which the task complexity can be switched according to the acoustic conditions and the application that is running. We have implemented a multiple mode policy outlined below:

Mode	Robot	Speaker position	Application
3	idle	frontal $\leq$ 50cm	spontaneous dialogue
2	idle	unknown	situated dialogue
1	motion	unknown	isolated commands

Spontaneous dialogue refers to unconstraint large vocabulary dialogue; whereas situated dialogue, refers to scripted dialogue dependent on the current situation. Here regular grammars can be used.

By switching between these three recognition modes the robot can demonstrate complex dialogue understanding (albeit in limited situations) while working robustly in adverse conditions (albeit with a limited vocabulary and syntax). For mode 3, the position of the speaker can be visually confirmed. If confirmation fails the robot goes back to a lower recognition mode.

To optimize performance it would be advantageous to have separate recognition engines optimized for each task and environment. For example for small vocabulary recognition such as in mode 1, word models (instead of phonemic models) would be preferable. Also one might want to use different acoustic models trained in matched conditions for the respective acoustic environment for each mode.

We decided not to do so for three reasons: Firstly, we wanted to keep the recognition platform as general as possible. Secondly, dealing with multiple acoustic models (AMs) and other recognition resources would have required more memory and/or latencies when task switches occur. Thirdly, we prefered a unified framework in which multiple tasks (based on the same recognition resources) can be executed simultaneously. This is because it is sometimes not possible to know in advance which of the three modes should be applied. (For example, for technical reasons, it cannot be determined in advance, whether the robot will remain stationary during the utterance)

Therefore we adopted a single engine, called Arthur, capable of performing fast and accurate large vocabulary spontaneous continuous speech recognition, as well as recognition based on regular grammars or isolated word recognition in a noisy environment. Likewise the AM was also designed to operate in all three modes. However for reasons of conciseness, we only report on results for recognition mode 3.

name	style	Description	f	m	utter
nws01	read	News articles	5	5	900
nws02	read	News articles	4	4	720
htl14	read	Travel conversations	8	8	2792
htl02	read	Travel conversations	8	8	2000
htl04	read	Travel conversations	6	6	3000
rwc01	spont.	Person to person convers.	8	8	80
eli01	spont.	Chatting to robot	5	5	297

Table 2: Data bases used for evaluation, indicating number of female and male speaker and total number of utterances.

	rwc01		htl02	
acoustic model	Corr	Acc	Corr	Acc
IPA9904	56.47	31.47	73.24	64.46
Sony0204	59.31	36.10	72.45	64.34

Table 3: Syllable recognition results on two data bases using acoustic models trained on read speech and on a combination of read and spontaneous speech.

4. Acoustic model

When speech is recorded with a far-field microphone two types of distortions affect the speech signal: room and actuator noises and reverberations. Both have a devastating effect on the recognition rate, when the recognizer is trained on clean speech [3]. A number of methods have been proposed to address this problem. For additive noise, the noise statistics can be estimated and subtracted in the spectral domain, to map the features closer to those of undistorted speech [1], alternatively the HMM model can be adapted online using an adaptation scheme such as MLLR [7] or model combination [5]. Finally training conditions can be matched by either training on data recorded in the same condition as the eventual application or by artificially creating such data through the addition of a noise signal.

The adaptive approaches suffer from having to estimate the spectral envelope of the noise and the signal to noise ratio, which can be erratic in non-stationary noise conditions. We adopted non-linear spectral subtraction as a simple feature space adaptation technique for stationary noise and otherwise opted to build static robust models suitable for a wide range of conditions.

In what follows the following parameters were used: 16kHz sampling frequency, 25ms frame length, 10ms frame shift, 25 dimensional feature vectors (12 MFCC coefficients + delta coefficients + delta C_0), 3 state HMM, no skip transitions, 1000 tied-state triphone HMMs with 16 diagonal Gaussian components. Spectral subtraction was used during training.

4.1. Training on spontaneous data

There are currently few databases that contain sufficient spontaneous speech for training of acoustic models. Part of the problem is, that soliciting truly spontaneous data in a recording scenario is difficult and expensive. We used a combination of read speech, solicited spontaneous speech and spontaneous speech from TV talk shows. The read speech consisted of 600 phonetically balanced sentences (sn601) as well as a collection of stories, and magazine articles (dra01). For the solicited spontaneous speech we used the Japanese portion of the VerbMobil data [6]. The TV data was obtained from spoken dialogues in talk shows which were segmented into individual utterances. We refer to the HMM trained on these corpora as "Sony0204". For our baseline we

dist to mic	noise	SNR	clean HMM	noise/rev HMM
Clean test data (htl04)				
–	–	–	65.28	61.3
Reverberated, noise-added test data (htl04)				
100cm	idling, fan off	17db	28.6	41.1
	idling, fan on	12db	23.0	36.5
	nodding 0.5Hz	20db	28.6	40.3
	walk on carpet	6db	13.2	21.6
300cm	idling, fan off	17db	25.8	36.8
	idling, fan on	12db	20.0	32.3
	nodding 0.5Hz	20db	25.5	36.1
	walk on carpet	6db	12.7	20.5
Test data recorded on robot (spontaneous utterances)				
100cm	power off		32.8	44.2
	+ idling noise		32.8	44.2
	power on, idling		31.7	39.7

Table 4: Syllable recognition rates for environmentally robust training for Sony0204

trained an HMM on the material released by the IPA community in April 1999 [9]. This is refered to as "IPA9904". Table 1 summarizes the statistics of the recordings.

We tested the spontaneous AM by performing continuous syllable recognition on two of our evaluation data bases (Table 2): htl02 a collection of read travel conversation and rwc01 a spontaneous corpus released by the real word computing consortium [10]. The experiments were done without dictionary or language model to isolate the effect of the acoustic model. Table 3 shows the syllable correctness and accuracy. The new model, trained on a mixture of read and spontaneous speech improved performance on the spontaneous data. Performance on the read speech corpus suffered, but only slightly so. More comprehensive results are shown in section 6.

4.2. Reverberation and noise training

In order to obtain environmental robustness we used the simple technique of training on a variety of environments to obtain a single "broadband" acoustic model. Environments were generated from clean speech by mixing in convolutional and additive noise. For convolutional noise we measured the impulse response of several rooms with varying distance to the recording microphone using the stretched pulse technique [2]. For additive noise we recorded various robot noises that are generated when the robot is idle or in motion. Training data was generated by first creating reverberated speech data and then adding the various noises.

To test the performance of the new HMM we conducted continuous Japanese syllable recognition experiments. We did not employ any dictionary or language model, in order to isolate the effect of the HMM from the other resources. Tests were performed on a clean read-speech travel dialogue database (htl04), on the same database with various additive and convolutional noises added, and on real data recorded on the robot. Care was taken that the noises reverberations were independent of those used in training. Table 4 summarizes the results. Apart from the clean test case, the noisy HMM outperformed the clean HMM by a large margin.

5. Dictionary and Language Model

Dictionary and Language model are generally task dependent and the applications running on the SDR-4X can supply their own to provide optimal coverage and recognition accuracy for a given task. However for unrestricted chat applications a "broad coverage" language model with a 20k word vocabulary is provided. This is for two reasons: (1) Because the LM stays resident latencies during task switches can be avoided. (2) Application developers are relieved from supplying their own LM for their applications.

To provide broad coverage, a spontaneous "chat" domain was chosen, which should impose minimal constraints on what the user is allowed to say.

Transcriptions of spontaneous speech are difficult to obtain in large enough quantities for training language models. Therefore we opted for collecting web pages. We had previously built an LM based on online novels and diaries as well as newspaper articles but this did not seem appropriate for our spontaneous chat application. We therefore collected web pages with a more spontaneous content by searching for the appearance of certain colloquial Japanese expressions which indicate an informal colloquial writing style.

The data collection method was tested on htl14 and rwc01. Two LMs were built: LM2 containing 1.9M sentences and LM3 containing 9.7M sentences. This was compared to our baseline (LM1) constructed from 8.4M newspaper sentences and 2.5M sentences taken from online diaries and novels on the basis of test-set perplexity (P.P.) and word accuracy (Acc) using the read-speech HMM (IPA9904). The result is summarized in the following table (OOV = out-of-vocabulary rate):

Task:	rwc01			htl14		
	Vocab	Acc	P.P.	OOV	Acc	P.P.	OOV
LM1	61K	39.3	136.8	0.50%	81.0	72.6	0.35%
LM2	64K	41.2	89.9	0.50%	79.5	58.9	0.56%
LM3	64K	44.0	82.6	0.50%	84.6	51.6	0.28%

It can be seen that the LMs obtained from the new collection method reduced the test-set perplexity of both the read speech as well as the spontaneous task. Although for LM2 the recognition accuracy drops for the hotel domain, LM3 is superior to the baseline for both tasks.

The result also shows the old wisdom that "bigger is better". However for our embedded application an LM was required that could be kept resident in memory to provide fast recognition and fast task switching. By using data compression, relatively high cutoff thresholds for both bi- and trigrams and reducing the vocabulary size to 20000 words, we succeeded in reducing the LM size to just above 5Mbyte with only a minor degradation in accuracy. By screening the LM training data and dictionary we succeeded to further improve accuracy and speed. To screen the LM training data, an LM was built from transcriptions of spontaneous data recorded from TV shows. This LM was used to measure the test-set perplexity of the various web pages used in building LM3. A fixed fraction of the this material with the highest test-set perplexity was then removed from the training data and the remaining web pages were used to build the spontaneous LM (LM4). We achieved best results by removing 20% of the training material in this way. The results could be improved further by screening the dictionary for single syllabic character (hiragana and katakana) words, which had resulted from our dictionary generation process, but were found to do more harm than good. Experimental results on the screening procedures are given in the following section.

	A	B	C	D	E	F	G
HMM	IPA9904	IPA9904	Sony0204	Sony0204	Sony0204	Sony0204	Sony0204
training condition	clean	clean	clean	noise/rev	noise/rev2	noise/rev2	noise/rev2
LM	LM1	LM3	LM3	LM3	LM3	LM4	LM4
Cut-off	0-1	0-1	0-1	0-1	0-1	10-10	10-10
Dictionary size	61k	64k	64k	64k	64k	20k Scrnd	20k Scrnd
Test Corpus (rec mic)	Word accuracy @ real time factor for clean data						
nws01 (c-38b)	95.01	81.21	79.74@0.92	76.53@1.12	77.73@0.92	72.07@0.81	72.07@0.68
nws02 (hmd25-1)	93.32	75.19	74.60@1.04	68.33@1.20	69.97@1.01	64.76@0.90	64.76@0.70
htl14 (c-38b)	84.93	83.64	81.69@0.92	82.01@1.12	81.64@0.90	82.97@0.79	82.97@0.65
eli01 (c-38b)	62.16	65.12	66.06 —	64.83 —	67.92 —	66.99 —	66.99 —
rwc01	40.87	43.22	52.14 —	50.72 —	54.06 —	53.36 —	53.36 —
Noise/Mic dist/SNR	Word accuracy for htl14 with added noise and reverberation						
idling / 50cm/20db	71.63	73.51	72.47	79.39	78.00	79.61	79.61
nodding/ 50cm/19db	68.52	70.67	69.25	77.51	75.35	77.52	77.52
idling /100cm/17db	60.75	63.91	61.30	74.87	72.63	75.50	75.50
nodding/100cm/16db	55.50	57.75	57.12	71.22	67.92	70.99	70.99

Table 5: Overall recognition results on a number of read and spontaneous corpora in clean and noisy conditions

6. Overall results

Our overall development process is summarized in table 5. We started with an HMM trained on read speech only, in clean conditions and a language model based on a combination of newspaper articles and novels (A). We achieved high accuracy on a clean news corpus, but relatively poor performance on spontaneous data. The LM build from Web pages chosen for their colloquial style improved performance on the spontaneous corpora, while the hotel corpus remained about the same and the news corpora suffered (B). When we switched to the spontaneous HMM trained from a mixture of read and spontaneous data (Sony0204) performance improved further on the spontaneous data, while the read-speech performance deteriorated slightly (C). Employing noise and reverberation training, results improved significantly on small vocabulary noisy test corpora for the recognition modes 1 and 2 (results not shown). Results for recognition mode 3 (distance to mic 50cm, robot idle or nodding) also improved (column D, bottom part of the table). For clean LVCSR tasks a small degradation for both spontaneous and read speech was observed. More significantly, real-time performance was affected. The reported real-time factors were measured in our simulation environment on a Sun Workstation at 300MHz. Previous experience had shown that real time factors are worse by about a factor of 2 on the embedded platform of the SDR-4X. For a real system real-time factors of more than 2 are unacceptable. To improve the real-time performance, we increased the amount of clean training data in the training of the reverberated HMM. This was simply done by re-using the same material repeatedly, i.e. weighting it more strongly, relative to the noisy/reverberated speech data. Optimal results were obtained when the weighting of clean to noisy speech data was about 1:1 (noise/rev2). The results are shown in column E. The real-time behavior improved. At the same time performance on clean speech improved, while the performance on noisy speech only suffered slightly. The result of screening the LM and Dictionary and compacting the LM using higher cut-offs is shown in column F. The main gain here was the reduction in LM size from about 50Mbytes to about 5Mbytes so that it could be kept resident for fast decoding and task switches. Finally, algorithmic changes to our recognition engine Arthur further improved the real-time performance (G).

7. Discussion

This contribution summarized the main developmental steps in building an environmentally robust large vocabulary continuous spontaneous speech recognition system for an entertainment robot. Even though the robot is exposed to a wide variety of environments, speakers and speaking styles, we followed a broad-band instead of an adaptive approach, by providing a single HMM and LM to cover this wide range of conditions. By carefully balancing different conditions we succeeded in building a robust system that runs in near real time on an embedded platform. It can be debated whether 60+% recognition accuracy for spontaneous speech is sufficient. Clearly a higher figure is desirable. However, for an entertainment application recognition errors are perhaps less critical than they are in task-oriented applications in which rate of, and time for, task completion are important. In addition to further improving accuracy, detecting, acknowledging and handling these mis-recognitions in a natural way within the entertainment dialogue strategy becomes an important research item.

8. References

[1] A. Acero. Acoustical and Environmental Robustness in Automatic Speech Recognition. PhD thesis, CMU, Sept. 1990.

[2] N. Aoshima. Computer-generated pulse signal applied for sound measurement. J. Acoust. Soc. Am., 69(5):1119–1123, Feb. 1981.

[3] J. C.Junqua and J. P. Haton. Robustness in automatic speech recognition. Kluwer Academic Publishers, 1996.

[4] C. Culhane. Session 7 - conversational and multi-lingual speech recognition. In Proc. 1996 DARPA Speech Recognition Workshop, pages 143–144, 1996.

[5] M. Gales and S. Young. Robust speech recognition in additive and convolutional noise using parallel model combination. IEEE. Trans on Speech and Audio Proc., 9:289–307, Jan. 1995.

[6] A. Kurematsu et al. Development of data collection and transliteration of japanese spontaneous database in the travel arrangement task domain. In Proc. Oriental COCOSDA, 1999.

[7] C. J. Leggetter and P. C. Woodland. Maximum likelihood linear regression for speaker adaptation of continuous density hidden markov models. Computer Speech and Language, 9:242–245, 1995.

[8] M. Padmanabhan and M. Picheny. Towards super-human speech recognition. In ISCA Turorial and Research Workshop, 2000.

[9] A. Shikano et al. Volunteer-based IPA Japanese dictation free software project. In Proc. Oriental COCOSDA, 1999.

[10] K. Tanaka et al. Design and data collection for a spoken dialogue database in the real world computint (rwc) program. Inf. Proc. Soc. of Japan SIG Notes, (11–7), May 1996.