# Characterizing and Processing Robot-Directed Speech

Paulina Varchavskaia, Paul Fitzpatrick, Cynthia Breazeal

AI Lab, MIT, Cambridge, USA
[paulina,paulfitz,cynthia]@ai.mit.edu

**Abstract.** Speech directed at infants and pets has properties that distinguish it from speech among adults [6]. Some of those properties are potentially useful for language learning. By careful design of form and behavior, robots can hope to evoke a similar speech register and take advantage of these properties. We report some preliminary data to support this claim, based on experiments carried out with the infant-like robot Kismet [4]. We then show how we can build a language model around an initial vocabulary, perhaps acquired from "cooperative" speech, and bootstrap from it to identify further candidate vocabulary items drawn from arbitrary speech in an unsupervised manner. We show how to cast this process in a form that can be largely implemented using a conventional speech recognition system [8], even though such systems are designed with very different applications in mind. This is advantageous since, after decades of research, such systems are expert at making acoustic judgments in a probabilistically sound way from acoustic, phonological, and language models.

**Keywords:** speech recognition, language learning, infant-directed speech, robot-directed speech, referential mapping, word-spotting.

## 1 Introduction

A natural-language interface is a desirable component of a humanoid robot. In the ideal, it allows for natural hands-free communication with the robot without necessitating any special skills on the human user's part. In practice, we must trade off flexibility of the interface with its robustness. Contemporary speech understanding systems rely on strong domain constraints to achieve high recognition accuracy [20]. This paper makes an initial exploration of how ASR techniques may be applied to the domain of robot-directed speech with flexibility that matches the expectations raised by the robot's humanoid form.

A crucial factor for the suitability of current speech recognition technology to a domain is the expected perplexity of sentences drawn from that domain. Perplexity is a measure of the average branching factor within the space of possible word sequences, and so generally grows with the size of the vocabulary. For example, the basic vocabulary used for most weather-related queries may be quite small, whereas for dictation it may be much larger and with a much less constrained grammar. In the first case speech recognition can be applied successfully for a large user population across noisy telephone lines [19], whereas in the second a good quality headset and extensive user training are required in practice. It is important to determine where robot-directed speech lies in this spectrum. This will depend on the nature of the task to which the robot is being applied, and the character of the robot itself. For this paper, we will consider the case of Kismet [4], an "infant-like" robot whose form and behavior is designed to elicit nurturing responses from humans.

We first look at approaches to speech interfaces taken by other groups. Then we briefly review the potential advantages of eliciting infant-directed speech. In Section 4 we present some preliminary results characterizing the nature of speech directed at Kismet. The remainder of the paper develops an unsupervised procedure for vocabulary extension and language modeling.

## 2 Background and motivation

Recent developments in speech research on robotic platforms have followed two basic approaches. The first approach builds on techniques developed for command-and-control style interfaces. These systems employ the standard strategy found in ASR research of limiting the recognizable vocabulary to a particular predetermined domain or task. For instance, the ROBITA robot [14] interprets command utterances and queries related to its function and creators, using a fixed vocabulary of 1,000 words. Within a small fixed domain fast performance with few errors becomes possible, at the expense of any ability to interpret out-of-domain utterances. But in many cases this is perfectly acceptable, since there is no sensible response available for such utterances even if they were modeled.

A second approach adopted by some roboticists [17, 15] is to allow adjustable (mainly growing) vocabularies. This introduces a great deal of complexity, but has the potential to lead to more open, general-purpose systems. Vocabulary extension is achieved through a label acquisition mechanism using either supervised or unsupervised learning algorithms. This approach was taken in CELL [17], Cross-channel Early Lexical Learning, where a robotic platform called Toco the Toucan was developed to implement a model of early human language acquisition. CELL is embodied by an active vision camera placed on a four degree of freedom motorized arm and augmented with expressive features to make it appear like a parrot. The system acquires lexical units from the following scenario: a human teacher places an object in front of the robot and describes it. The visual system extracts color and shape properties of the object, and CELL learns on-line a lexicon of color and shape terms grounded in the representations of objects. The terms learned need not pertain to color or shape exclusively - CELL has the potential to learn any words, the problem being that of deciding which lexical items to associate with which semantic categories. In CELL, associations between linguistic and contextual channels are chosen on the basis of maximum mutual information. Also in [15], a Pioneer-1 mobile robot was programmed with a system to cluster its sensory experiences using an unsupervised learning algorithm. In this way the robot extends its vocabulary by associating sets of sensory features with the spoken labels that are most frequently uttered in their presence.

We share the goal of automatically acquiring new vocabulary. We are particularly interested in augmenting unsupervised techniques that work on arbitrary speech with more specialized methods that make use of human cooperation. Section 3 looks at infant-directed speech, a special speech register which many claim has interesting properties for facilitating language learning. A similar register arguably exists for "pet-directed speech" [6], so we hope that an infant-like robot may also evoke speech with similar properties. Section 4 presents some preliminary results to verify whether this is the case. The remainder of the paper shows how to build a language model around any vocabulary we can extract in this way, and use that model to locate further candidates for vocabulary extension.

## 3 Infant-directed speech

When interacting with a youthful-appearing robot such as Kismet, we can expect that the speech input may have specialized characteristics similar to those of infant-directed speech (IDS). This section examines some of the properties of IDS so they may inform our expectations of the nature of Kismet-directed speech. We examined the following two questions regarding the nature of IDS:

- Does it include a substantial proportion of single-word utterances? Presenting words in isolation side-steps the problematic issue of word segmentation.
- How often, if at all, is it clearly enunciated and slowed down compared to normal speech? Overarticulated speech may be helpful to infants, but has important consequences for artificial speech recognizers.

**Isolated words** Whether isolated words in parental speech help infants learn has been a matter of some debate. It has been shown that infant-directed utterances are usually short with longer pauses between words (e.g., research cited in [18]), but also that they do not necessarily contain a significant proportion of isolated words [1]. Another study [5] presents evidence that isolated words are in fact a reliable feature of infant-directed speech, and that infants' early word acquisition may be facilitated by their presence. In particular, the authors find that the frequency of exposure to a word in isolation is a better predictor of whether the word will be learned, than the total frequency of exposure. This suggests that isolated words may be easier for infants to process and learn. Equally importantly for us, however, is the evidence for a substantial presence of isolated words in IDS: 9% found in [5] and 20% reported in [1]. If Kismet achieves its purpose of eliciting nurturing behavior from humans, then perhaps we can expect a similar proportion of Kismet-directed speech to consist of single-word utterances. This hypothesis will undergo a preliminary evaluation in Section 4.

**Enunciated speech and "vocal shaping"** The tendency of humans to slow down and overarticulate their utterances when they meet with misunderstanding has been reported as a problem in the ASR community [12]. Such enunciated speech degrades considerably the performance of speech recognition systems which were trained on natural speech only. If we find that human caretakers tend to address Kismet with overarticulated speech, its presence becomes an important issue to be addressed by the robot's perceptual system.

In infant-directed speech, we might expect overarticulation to occur in an instructional context, when a caretaker deliberately introduces the infant to a new word or corrects a mispronunciation. Another possible strategy is that of "shaping" of the infant's pronunciation by selecting and repeating the mispronounced part of the word until a satisfactory result is reached. There is evidence that parents may employ such a strategy, but it appears to be mostly at the anecdotal level.

## 4 Exploring robot-directed speech

This section describes a preliminary study of interactions between young children and the Kismet robot in the context of teaching the robot new words. During these sessions, the robot was engaging in proto-conversational turn-taking, where its responses to utterances of the children were random affective babble. A very minimal mechanism for vocal mimicry and vocabulary extension was present. The purpose of our study is to identify ways to improve the speech interface on the robot based on a better knowledge of the properties of speech directed at this particular robot.

### 4.1 Robot configuration

During these experiments the robot was engaging in proto-conversational turn-taking as described in [4], augmented with the following command-and-control style grammar. Sentences that began with phrases such as "say", "can you say", "try" etc. were treated as requests for the robot to repeat the phonetic sequence that followed them. If, after the robot repeated a sequence, a positive phrase such as "yes" or "good robot" was heard, the sequence would be entered in the vocabulary. If not, no action was taken unless the human's next utterance was similar to the first, in which case it was assumed to be a correction and the robot would repeat it. Because of the relatively low accuracy of phoneme-level recognition, such corrections are the rule rather than the exception.

### 4.2 Data collection

For this preliminary study, we drew on recordings originally made for Sherry Turkle's research on children's perception of technology and identity. We analyzed video of 13 children aged from 5 to 10 years old interacting with the robot. Each session lasted approximately 20 minutes. In two of the sessions, two children are playing with the robot at the same time. In the rest of the sessions, only one child is present with the robot.

### 4.3 Preliminary data analysis

We were interested in determining whether any of the following strategies are present in Kismet-directed speech:

– single-word utterances (words spoken in isolation)
– enunciated speech
– vocal shaping (partial, directed corrections)
– vocal mimicry of Kismet's babble

A total of 831 utterances were transcribed from the 13 sessions of children playing with the robot. We observed a wide variation of strategies among subjects. The following preliminary results include a measure of standard deviations, which are mentioned to give an idea of the wide range of the data, and should not be read to imply that the data follows a Gaussian distribution. The total number of utterances varied from subject to subject in the range between 19 and 169, with a mean of 64 (standard deviation of 44, based on a sample of 13) utterances per subject.

**Isolated words** These are fairly common; 303 utterances, or 36.5% consisted of a single word said in isolation. The percentage of single-word utterances had a distribution among subjects with a mean at 34.8 and a deviation of 21.1. Even when we exclude both greetings and the robot's name from counts of single-word utterances, we get a distribution centered around 20.3% with a standard deviation of 18.5%. This still accounts for a substantial proportion of all recorded Kismet-directed speech. However, almost half the subjects use less than 10% isolated words, even in this teaching context (see Table 1).

**Enunciated speech** Also common is enunciated speech; 27.4% of the transcribed utterances (228) contained enunciated speech. An utterance was counted as "enunciated speech" whenever deliberate pauses between words or syllables within a word, and vowel lengthening were used. The count therefore includes the very frequent examples where a subject would ask the robot to repeat a word, e.g. "Kismet, can you say: GREEN?". In such examples, GREEN would be the only enunciated part of the utterance but the whole question was counted as containing enunciated speech. The mean proportion of enunciated speech is 25.6% with a deviation of 20.4%, which again shows a large variation.

**Vocal shaping** In the whole body of data we have discovered only 6 plausible instances (0.7%) of vocal shaping. It may not be an important teaching strategy, or it may not be evoked by a mimicry system that is not responding reliably enough to the teacher.

**Vocal mimicry** There were 23 cases of children imitating the babbling sounds that Kismet made, which accounts for 2.8% of the transcribed utterances. However, most children did not use this strategy at all.

### 4.4 Discussion

Qualitatively, the results presented above seem encouraging. However, before we draw any conclusions from the analysis, we must realize that in this instance, the process of gathering the data and the method of analysis had several shortcomings. The data itself, as was mentioned earlier, came from recordings of interactions set up for the purposes of an unrelated sociological study of children, a collaboration effort between the AI Lab and the MIT Initiative on Technology and Self. The interaction sessions were not set up as controlled experiments, and do not necessarily represent spontaneous Kismet-directed speech. In particular, on all occasions but one, at some point during the interaction, children were instructed to make use of the currently implemented command-and-control system to get the robot to repeat words after them. In some cases, once that happened, the subject was so concerned with getting the robot to repeat a word that anything else simply disappeared from the interaction. On three occasions, the subjects were instructed to use the "say" keyword as soon as they sat in front of the robot. When subjects are so clearly focused on a teaching scenario, we can expect the

| subject | # utterances | # single-word utterances | % | # single-word greetings | # kismet utterances | % without greetings, kismet | enunciated | % |
|---|---|---|---|---|---|---|---|---|
| 1 | 94 | 65 | 69.2 | 0 | 30 | 37.2 | 14 | 14.9 |
| 2 | 19 | 9 | 47.4 | 1 | 2 | 31.6 | 5 | 26.3 |
| 3 | 128 | 69 | 54.0 | 11 | 46 | 9.3 | 19 | 14.8 |
| 4 | 37 | 17 | 46.0 | 2 | 7 | 21.6 | 1 | 2.7 |
| 5 | 26 | 9 | 34.7 | 3 | 0 | 23.1 | 0 | 0.0 |
| 6 | 61 | 14 | 23.0 | 9 | 0 | 8.2 | 15 | 24.6 |
| 7 | 34 | 2 | 5.9 | 1 | 0 | 2.9 | 5 | 14.7 |
| 8 | 73 | 43 | 58.9 | 0 | 0 | 58.9 | 20 | 27.4 |
| 9 | 169 | 39 | 23.1 | 8 | 9 | 13.0 | 63 | 37.3 |
| 10 | 32 | 17 | 53.1 | 0 | 2 | 46.9 | 21 | 65.6 |
| 11 | 56 | 7 | 12.5 | 3 | 1 | 5.4 | 22 | 39.2 |
| 12 | 33 | 5 | 15.2 | 5 | 0 | 0.0 | 2 | 6.1 |
| 13 | 69 | 7 | 10.1 | 3 | 0 | 5.8 | 41 | 59.4 |
| **total** | 831 | 303 | | 46 | 97 | | 228 | |
| **mean** | | | 34.8 | | | 20.3 | | 25.6 |
| **deviation** | | | 21.1 | | | 18.5 | | 20.4 |

**Table 1.** Analysis of Kismet-directed speech

proportion of isolated words, for instance, to be unnaturally high.

Note also that as of now, we have no measure of accuracy of the transcriptions, which were done by hand by one transcriber, from audio that sometimes had poor quality. Given the focus of the analysis, only Kismet-directed speech was noted from each interaction, excluding any conversations that the child may have had with other humans who were present during the session. Deciding which utterances to transcribe was clearly another judgment call that we cannot validate here yet. Finally, since the speech was transcribed by hand, we cannot claim a scientific definition of an utterance (e.g., by pause duration) but must rely on one person's judgement call again.

However, this preliminary analysis shows promise in that we have found many instances of isolated words in Kismet-directed speech, suggesting that Kismet's environment may indeed be scaffolded for word learning. However, fluent speech is still prevalent even in a teaching scenario, and so an unsupervised learning algorithm will be needed to find new words in this case. We have also found that a substantial proportion of speech was enunciated. Counter-intuitively such speech can present problems for the speech recognizer, but at the same time opens new possibilities. For an improved word-learning interface, it may be possible to discriminate between natural and enunciated speech to detect instances of pronunciation teaching (this approach was taken in the ASR community, for example in [12]). On the other hand, the strategy of vocal shaping was not clearly present in the interactions, and there were few cases of mimicry.

Having completed this exploratory study, we now plan to follow up the results with more tightly controlled experiments specifically designed to elucidate the nature of the speech input to the robot.

## 5 Unsupervised vocabulary extension

This section develops a technique to bootstrap from an initial vocabulary (perhaps introduced by the methods described in Section 4) by building an explicit model of unrecognized parts of utterances. The purpose of this background model is both to improve recognition accuracy on the initial vocabulary and to automatically identify candidates for vocabulary extension. This work draws on research in word spotting and speech recognition. We will bootstrap from a minimal background model, similar to that used in word-spotting, to a much stronger model where many more word or phrase clusters have been "moved to the foreground" and explicitly modeled. This is intended both to boost performance on the original vocabulary by increasing the effectiveness of the language model, and to identify candidates for automatic vocabulary extension.

The remainder of this section shows how a conventional speech recognizer can be convinced to cluster frequently occurring acoustic patterns, without requiring the existence of transcribed data.

**Clustering algorithm** A speech recognizer with a phone-based "OOV" (out-of-vocabulary) model is able to recover an approximate phonetic representation for words or word sequences that are not in its vocabulary. If commonly occurring phone sequences can be located, then adding them to the vocabulary will allow the language model to capture their co-occurrence with words in the original vocabulary, potentially boosting recognition performance. This

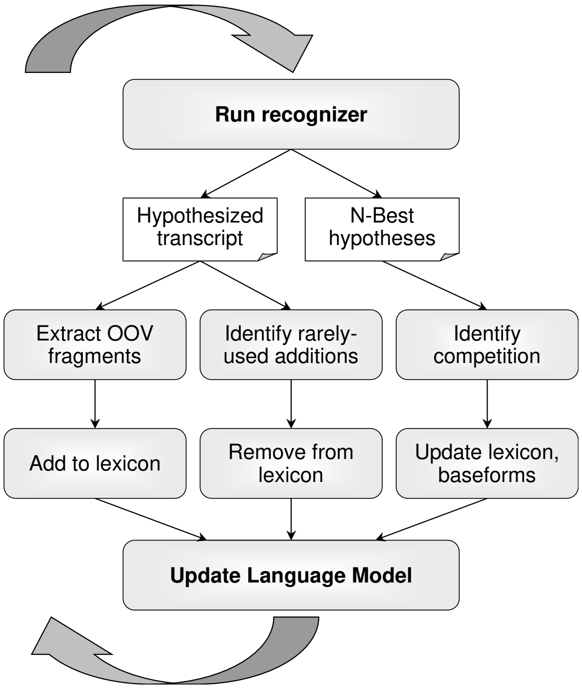


**Fig. 1.** The iterative clustering procedure.

suggests building a "clustering engine" that scans the output of the speech recognizer, correlates OOV phonetic sequences across all the utterances, and updates the vocabulary with any frequent, robust phone sequences it finds. While this is feasible, the kind of judgments the clustering engine needs to make about acoustic similarity and alignment are exactly those at which the speech recognizer is most adept.

The clustering procedure we adopted is shown in Figure 1. An *n*gram-based language model is initialized uniformly. Unrecognized words are explicitly represented using a phone-based OOV model, described in the next section. The recognizer is then run on a large set of untranscribed data. The phonetic and word level outputs of the recognizer are compared so that occurrences of OOV fragments can be assigned a phonetic transcription. A randomly cropped subset of these are tentatively entered into the vocabulary, without any attempt yet to evaluate their significance (e.g. whether they occur frequently, whether they are similar to existing vocabulary, etc.). The hypotheses made by the recognizer are used to retrain the language model, making sure to give the new additions some probability in the model. Then the recognizer runs using the new language model and the process iterates. The recognizer's output can be used to evaluate the worth of the new "vocabulary" entries. The following sections detail how to eliminate vocabulary items the recognizer finds little use for, and how to detect and resolve competition between similar items.

**Extracting OOV phone sequences** We use the speech recognizer system developed by the SLS group at MIT [8]. The recognizer is augmented with the OOV model developed by Bazzi in [2]. This model can match an arbitrary sequence of phones, and has a phone bigram to capture phonotactic constraints. The OOV model is placed in parallel with the models for the words in the vocabulary. A cost parameter can control how much the OOV model is used at the expense of the in-vocabulary models. This value was fixed at zero throughout the experiments described in this paper, since it was more convenient to control usage at the level of the language model. The bigram used in this project is exactly the one used in [2], with no training for the particular domain.

Phone sequences are translated to phonemes, then inserted as new entries in the recognizer's lexicon.

**Dealing with rarely-used additions** If a phoneme sequence introduced into the vocabulary is actually a common sound sequence in the acoustic data, then the recognizer will pick it up and use it in the next iteration. Otherwise, it just will not appear very often in hypotheses. After each iteration a histogram of phoneme sequence occurrences in the output of the recognizer is generated, and those below a threshold are cut.

**Dealing with competing additions** Very often, two or more very similar phoneme sequences will be added to the vocabulary. If the sounds they represent are in fact commonly occurring, both are likely to prosper and be used more or less interchangeably by the recognizer. This is unfortunate for language modeling purposes, since their statistics will not be pooled and so will be less robust. Happily, the output of the recognizer makes such situations very easy to detect. In particular, this kind of confusion can be uncovered through analysis of the N-best utterance hypotheses.

If we imaging aligning a set of N-best hypothesis sentences for a particular utterance, then competition is indicated if two vocabulary items exhibit both of these properties:

- ▷ Horizontally repulsive - if one of the items appears in a single hypothesis, the other will not appear in a nearby location within the same hypothesis
- ▷ Vertically attractive - the items frequently occur in the same location within different hypotheses

Since the utterances in this domain are generally short and simple, it did not prove necessary to rigorously align the hypotheses. Instead, items were considered to be aligned based simply on the vocabulary items preceding and succeeding them. It is important to measure both the attractive and repulsive conditions to distinguish competition from vocabulary items that are simply very likely to occur in close proximity.

Accumulating statistics about the above two properties across all utterances gives a reliable measure of whether two vocabulary items are essentially acoustically equivalent to the recognizer. If they are, they can be merged or pruned so that the statistics maintained by the language model will be well trained. For clear-cut cases, the competing items are merged as alternatives in the list of pronunciation variants for a single vocabulary unit. or one item is simply deleted, as appropriate.

Here is an example of this process in operation. In this example, "phone" is a keyword present in the initial vocabulary. These are the 10-best hypotheses for the given utterance:

"what is the phone number for victor zue"

```
<oov> phone (nahmber) (mihterz) (yuw)
<oov> phone (nahmber) (mihterz) (zyuw)
<oov> phone (nahmber) (mihterz) (uw)
<oov> phone (nahmber) (mihterz) (zuw)
<oov> phone (ahmberf) (mihterz) (zyuw)
<oov> phone (ahmberf) (mihterz) (yuw)
<oov> (axfaanah) (mberfaxr) (mihterz)
(zyuw)
<oov> (axfaanah) (mberfaxr) (mihterz) (yuw)
<oov> phone (ahmberf) (mihterz) (zuw)
<oov> phone (ahmberf) (mihterz) (uw)
```

The "<oov>" symbol corresponds to an out of vocabulary sequence. The sequences within parentheses are uses of items added to the vocabulary in a prior iteration of the algorithm. From this single utterance, we acquire evidence that:

- ▷ The entry for (ax f aa n ah) may be competing with the keyword "phone". If this holds up statistically across all the utterances, the entry will be destroyed.
- ▷ (n ah m b er), (m b er f axr) and (ah m b er f) may be competing. They are compared against each other because all of them are followed by the same sequence (m ih t er z) and many of them are preceded by the same word "phone".
- ▷ (y uw), (z y uw), and (uw) may be competing

All of these will be patched up for the next iteration. This use of the N-best utterance hypotheses is reminiscent of their application to computing a measure of recognition confidence in [11].

**Testing for convergence** For any iterative procedure, it is important to know when to stop. If we have a collection of transcribed utterances, we can track the keyword error rate on that data and halt when the increment in performance is sufficiently small. Keywords here refer to the initial vocabulary.

If there is no transcribed data, then we cannot directly measure the error rate. We can however bound the rate at which it is changing by comparing keyword locations in the output of the recognizer between iterations. If few keywords are shifting location, then the error rate cannot be changing above a certain bound. We can therefore place a convergence criterion on this bound rather than on the actual keyword error rate. It is important to just measure changes in keyword locations, and not changes in vocabulary items added by clustering.

## 6 Experiments in vocabulary extension

The unsupervised procedure described in the previous section is intended to both improve recognition accuracy on the initial vocabulary, and to identify candidates for vocabulary extension. This section describes experiments that demonstrate to what degree these goals were achieved. To facilitate comparison of this component with other ASR systems, results are quoted for a domain called LCSInfo [9] developed by the SLS group at MIT. This domain consists of queries about personnel – their addresses, phone numbers etc. Very preliminary results for Kismet-directed speech are also given.

### 6.1 Experiment 1: qualitative results

This section describes the candidate vocabulary discovered by the clustering procedure. Numerical, performance-related results are reported in the next section.

Results given here are from a clustering session with an initial vocabulary of five keywords (email, phone, room, office, address), run on a set of 1566 utterances. Transcriptions for the utterances were available for testing but were not used by the clustering procedure. Here are the top 10 clusters discovered on a very typical run, ranked by decreasing frequency of occurrence:

| | | | |
|---|---|---|---|
| 1 | n ah m b er | 6 | p l iy z |
| 2 | w eh r ih z | 7 | ae ng k y uw |
| 3 | w ah t ih z | 8 | n ow |
| 4 | t eh l m iy | 9 | hh aw ax b aw |
| 5 | k ix n y uw | 10 | g r uw p |

These clusters are used consistently by the recognizer in places corresponding to: "number, where_is, what_is, tell_me, can_you, please, thank_you, no, how_about, group," respectively in the transcription. The first, /n ah m b er/, is very frequent because of phrases like "phone number", "room number", and "office number". Once it appears as a cluster the language model is immediately able to improve recognition performance on those keywords.

Every now and then during clustering a "parasite" appears such as /dh ax f ow n/ (from an instance of "the phone" that the recognizer fails to spot) or /iy n eh l/ (from "email"). These have the potential

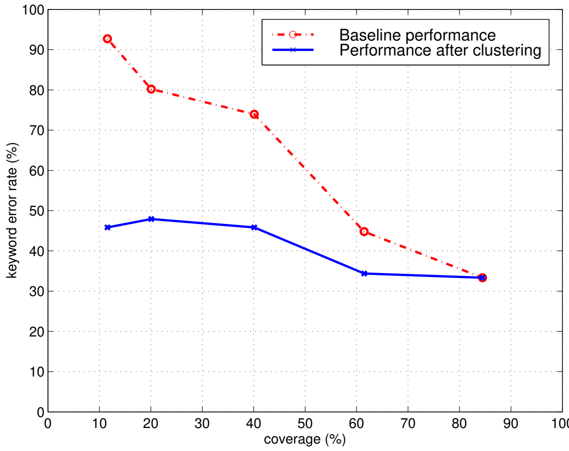


**Fig. 2.** Keyword error rate of baseline recognizer and clustering recognizer as total coverage varies.

to interfere with the detection of the keywords they resemble acoustically. But as soon as they have any success, they are detected and eliminated as described earlier. It is possible that if a parasite doesn't get greedy, and for example limits itself to one person's pronunciation of a keyword, that it will not be detected, although we didn't see any examples of this happening.

### 6.2 Experiment 2: quantitative results

For experiments involving small vocabularies, it is appropriate to measure performance in terms of Keyword Error Rate (KER). Here this is taken to be:

$$KER = \frac{F + M}{T} * 100 \qquad (1)$$

with:
F = Number of false or poorly localized detections
M = Number of missed detections
T = True number of keyword occurrences in data

A detection is only counted as such if it occurs at the right time. Specifically, the midpoint of the hypothesized time interval must lie within the true time interval the keyword occupies. We take forced alignments of the test set as ground truth. This means that for testing it is better to omit utterances with artifacts and words outside the full vocabulary, so that the forced alignment is likely to be sufficiently precise.

The experiments here are designed to identify when clustering leads to reduced error rates on a keyword vocabulary. Since the form of clustering addressed in this paper is fundamentally about extending the vocabulary, we would expect it to have little effect if the vocabulary is already large enough to give good coverage. We would expect it to offer the greatest improvement when the vocabulary is smallest. To measure the effect of coverage, a complete vocabulary for this domain was used, and then made smaller and smaller by incrementally removing the most infrequent words. A set of keywords were chosen and kept constant and in the vocabulary across all the experiments so the results would not be confounded by properties of the keywords themselves. The same set of keywords were used as in the previous section.

Clustering is again performed without making any use of transcripts. To truly eliminate any dependence on the transcripts, an acoustic model trained only on a different dataset was used. This reduced performance but made it easier to interpret the results.

Figure 2 shows a plot of error rates on the test data as the size of the vocabulary is varied to provide different degrees of coverage. The most striking result is that the clustering mechanism reduces the sensitivity of performance to drops in coverage. In this scenario, the error rate achieved with the full vocabulary (which gives 84.5% coverage on the training data) is 33.3%. When the coverage is low, the clustered solution error rate remains under 50% - in relative terms, the error increases by at most a half of its best value. Straight application of a language model gives error rates that more than double or treble the error rate.

As a reference point, the keyword error rate using a language model trained with the full vocabulary on the full set of transcriptions with an acoustic model trained on all available data gives an 8.3% KER.

### 6.3 Experiment 3: Kismet-directed speech

An experiment was carried out for data drawn from robot-directed speech collected for the Kismet robot. This data comes from an earlier series of recording sessions [7] rather than the ones described in Section 3. Early results are promising – semantically salient words such as “kismet”, “no”, “sorry”, “robot”, “okay” appear among the top ten clusters. But this work is in a very preliminary stage, since an acoustic model needs to be trained up for the robot's microphone configuration and environment.

## 7 Conclusions and Future Directions

The work described in this paper is not as yet a unified whole. We are approaching the question of language for a humanoid robot from several directions. One direction is concerned with characterizing and influencing the speech register used when people address the robot. Another examines how to extract vocabulary items from such speech, be it cooperative or otherwise. Other work, not described here, is addressing the crucial issue of binding vocabulary to meaning. One line of research under way is to use transient, task-dependent vocabularies to communicate the temporal structure of processes. Another line of research looks more generally at how a robot can establish a shared basis for communication with humans by learning expressive verbal behaviors as well as acquiring the humans' existing linguistic labels.

Parents tend to interpret their children's first utterances very generously and often attribute meaning and intent where there may be none [3]. It has been shown, however, that such a strategy may indeed help infants coordinate meaning and sound and learn to express themselves verbally. Pepperberg [16] formalized the concept into a teaching technique called referential mapping. The strategy is for the teacher to treat the pupil's spontaneous utterances as meaningful, and act upon them. This, it is shown, will encourage the pupil to associate the utterance with the meaning that the teacher originally gave it, so the student will use the same vocalization again in the future to make a similar request or statement. The technique was successfully used in aiding the development of children with special needs. In future work, we hope to apply this technique to build a shared basis for meaningful communication between the human and the robot.


## Acknowledgements

The authors would like to thank Sherry Turkle, Jen Audley, Anita Chan, Tamara Knutsen, Becky Hurwitz, and the MIT Initiative on Technology and Self, for making available the video recordings that were analyzed in this paper.

Parts of this work rely heavily on speech recognition tools and corpora developed by the SLS group at MIT.

Funds for this project were provided by DARPA as part of the "Natural Tasking of Robots Based on Human Interaction Cues" project under contract number DABT 63-00-C-10102, and by the Nippon Telegraph and Telephone Corporation as part of the NTT/MIT Collaboration Agreement.


## References


[1] R.N. Aslin, J.Z. Woodward, N.P. LaMendola, and T.G. Bever. Models of word segmentation in fluent maternal speech to infants. In J.L. Morgan and K. Demuth, editors, *Signal to Syntax: Bootstrapping From Speech to Grammar in Early Acquisition*. Lawrence Erlbaum Associates: Mahwah, NJ, 1996.

[2] I. Bazzi and J.R. Glass. Modeling out-of-vocabulary words for robust speech recognition. In *Proc. 6th International Conference on Spoken Language Processing*, Beijing, China, October 2000.

[3] P. Bloom. *How Children Learn the Meaning of Words*. Cambridge: MIT Press, 2000.

[4] C. Breazeal. *Sociable Machines: Expressive Social Exchange Between Humans and Robots*. PhD thesis, MIT Department of Electrical Engineering and Computer Science, 2000.

[5] M.R. Brent and J.M. Siskind. The role of exposure to isolated words in early vocabulary development. *Cognition*, 81:B33–B44, 2001.

[6] D. Burnham, E. Francis, C. Kitamura, U. Vollmer-Conna, V. Averkiou, A. Olley, and C. Paterson. Are you my little pussy-cat? acoustic, phonetic and affective qualities of infant- and pet-directed speech. In *Proc. 5th International Conference on Spoken Language Processing*, volume 2, pages 453–456, 1998.

[7] C.Breazeal and L. Aryananda. Recognition of affective communicative intent in robot-directed speech. In *Proceedings of Humanoids 2000*, Cambridge, MA, September 2000.

[8] J. Glass, J. Chang, and M. McCandless. A probabilistic framework for feature-based speech recognition. In *Proc. International Conference on Spoken Language Processing*, pages 2277–2280, 1996.

[9] J. Glass and E. Weinstein. Speechbuilder: Facilitating spoken dialogue systems development. In *7th European Conference on Speech Communication and Technology*, Aalborg, Denmark, September 2001.

[10] A.L. Gorin, D. Petrovksa-Delacrtaz, G. Riccardi, and J.H. Wright. Learning spoken language without transcriptions. In *Proc. IEEE Automatic Speech Recognition and Understanding Workshop*, Colorado, 1999.

[11] T.J. Hazen and I. Bazzi. A comparison and combination of methods for oov word detection and word confidence scoring. In *Proc. International Conference on Acoustics*, Salt Lake City, Utah, May 2001.

[12] J. Hirschberg, D. Litman, and M. Swerts. Prosodic cues to recognition errors. In *ASRU*, 1999.

[13] P. Jeanrenaud, K. Ng, M. Siu, J.R. Rohlicek, and H. Gish. Phonetic-based word spotter: Various configurations and application to event spotting. In *Proc. EUROSPEECH*, 1993.

[14] Y. Matsusaka and T. Kobayashi. Human interface of humanoid robot realizing group communication in real space. In *Proc. Second International Symposium on Humanoid Robots*, pages 188–193, 1999.

[15] T. Oates, Z. Eyler-Walker, and P. Cohen. Toward natural language interfaces for robotic agents: Grounding linguistic meaning in sensors. In *Proceedings of the 4th International Conference on Autonomous Agents*, pages 227–228, 2000.

[16] I. Pepperberg. Referential mapping: A technique for attaching functional significance to the innovative utterances of an african grey parrot. *Applied Psycholinguistics*, 11:23–44, 1990.

[17] D.K. Roy. *Learning Words from Sights and Sounds: A Computational Model*. PhD thesis, MIT, September 1999.

[18] J.F. Werker, V.L. Lloyd, J.E. Pegg, and L. Polka. Putting the baby in the bootstraps: Toward a more complete understanding of the role of the input in infant speech processing. In J.L. Morgan and K. Demuth, editors, *Signal to Syntax: Bootstrapping From Speech to Grammar in Early Acquisition*, pages 427–447. Lawrence Erlbaum Associates: Mahwah, NJ, 1996.

[19] V. Zue, J. Glass, J. Plifroni, C. Pao, and T.J. Hazen. Jupiter: A telephone-based conversation interface for weather information. *IEEE Transactions on Speech and Audio Processing*, 8:100–112, 2000.

[20] V. Zue and J.R. Glass. Conversational interfaces: Advances and challenges. *Proceedings of the IEEE, Special Issue on Spoken Language Processing*, Vol. 88, August 2000.