CHAPTER 9

First words: working with speech

The trouble with having an open mind, of course, is that people will insist on coming along and trying to put things in it. (Pratchett, 1989)

Speech sounds form a special category of percept, since speech is very much a cultural invention. Many of its properties are simply agreed to, rather than grounded in any immediate physical necessity There are of course many physical constraints on speech, but within that space there is huge potential diversity. And in fact, as a communication protocol, speech is very flexible. There are special schemes for talking to children, or pets. So it is quite easy to imagine that we could borrow one of these schemes for robots. One of the goals of the Kismet robot in our group was to evoke the "motherese" style of speech, for functional benefits (Varchavskaia et al., 2001). There are many robotics projects looking at various aspect of speech such as the development of vocabulary and/or grammar from various forms of experience (Roy and Pentland, 2002; Steels, 1996). The goal of this chapter is to produce a real-time system for extending vocabulary, augmented with a slower offline process for refinement, just as was the case for object recognition in Chapter 5.

9.1 The microphones

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. The first trade-off is the physical interface. For best results with contemporary speech recognition techniques, a high-quality microphone close to the mouth is desirable. On Kismet, a wireless clip-on or hand-held microphone was used, as shown in Figure 9-1. This caused some difficulties, because given Kismet's anthropomorphic face and prominent bright-pink ears, people expected the robot to be able to hear them directly without any intermediary. Unfortunately placing microphones in the ears or anywhere else in the head would have been completely useless, since all the motors controlling facial features were very noisy (and the ear motors were perhaps noisiest of all). On Cog, a microphone array was installed across its torso. This meant that a person interacting with the robot did not need to be instrumented, and natural human behavior when they want to be heard – speaking louder or coming closer – did in fact make the robot hear better. If background noise is high, there is an auxiliary wireless microphone which subsumes the microphone array.

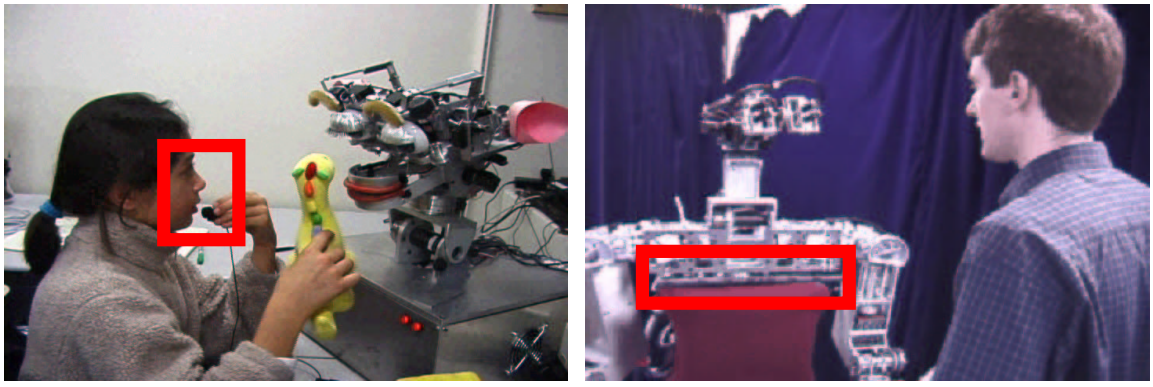

Figure 9-1: For Kismet, a clip-on/hand-held microphone was used (left). Anyone interacting with the robot needed to be informed of this. Experience showed that people would frequently forget to use the microphone if it was not clipped on – it was not an intuitive interface for face-to-face communication. On Cog, a commercial microphone array was installed (right), with a clip-on microphone available for when there was excessive background noise – Cog is located right beside a busy machine shop. This meant that the robot could always respond to voice in its vicinity, and if its responses were poor, the human could either move closer, speak louder (natural responses) or pick up the back-up microphone. The transition between microphones is handled automatically.

9.2 Infant-directed speech

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 (Zue et al., 2000), 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 presumably depend on the nature of the task to which the robot is being applied, and the character of the robot itself. We evaluated this for Kismet (Varchavskaia et al., 2001). When interacting with a youthful-appearing robot such as Kismet, our hope was that the speech input may have specialized characteristics similar to those of infant-directed speech (IDS). In particular, we were interested in the following:

- ▷ Does speech directed at Kismet 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 the speech clearly enunciated and slowed down compared to normal speech? Overarticulated speech may be helpful to infants, but can be challenging for artificial speech recognizers trained on normal speech.

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 (see for example Werker et al. (1996)), but also that they do not necessarily contain a significant proportion of isolated words (Aslin et al., 1996). Another study (Brent and Siskind, 2001) 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 Brent and Siskind (2001) and 20% reported in Aslin et al. (1996). If Kismet achieves its purpose of eliciting nurturing behavior from humans, then we would expect a similar proportion of Kismet-directed speech to consist of single-word utterances.

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 (Hirschberg et al., 1999). 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.

A study was made of interactions between young children and the Kismet robot in the context of teaching the robot new words. The sessions were organized by the MIT Initiative on Technology and Self. 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 the study was 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.

During these experiments the robot was engaging in proto-conversational turn-taking as described in Breazeal (2000), 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.

Video of 13 children aged from 5 to 10 years old interacting with the robot was analyzed. 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. 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.

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.

9.2.1 Discussion

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 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 Hirschberg et al. (1999)). On the other hand, the strategy of vocal shaping was not clearly present in the interactions, and there were few cases of mimicry.

9.3 Automatic language modeling

This section develops a technique to bootstrap from an initial vocabulary (distilled perhaps from isolated word utterances) 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.

9.3.1 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 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 adopted is shown in Figure 9-2. 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.

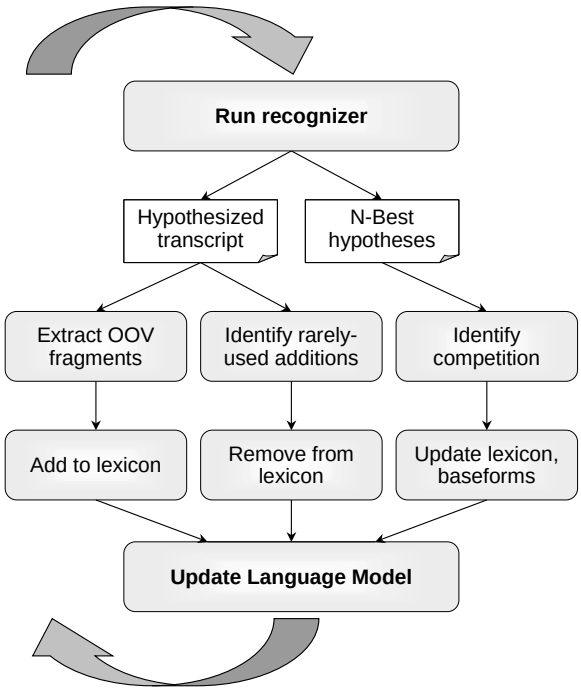

Figure 9-2: The iterative clustering procedure for segmenting speech. A conventional speech recognition system is used to evaluate how useful particular phoneme sequences are for describing the training data. Useful sequences are added to lexicon, otherwise they are dropped.

9.3.2 Extracting OOV phone sequences

The speech recognizer system developed by the Spoken Language Systems group at MIT was used (Glass et al., 1996). The recognizer is augmented with the OOV model developed by Bazzi and Glass (2000). 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 (Bazzi and Glass, 2000), with no training for the particular domain.

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

9.3.3 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.

9.3.4 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 (Hazen and Bazzi, 2001).

9.3.5 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.

9.4 Offline 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 (Glass and Weinstein, 2001) developed by the Spoken Language Systems 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.

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:

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

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

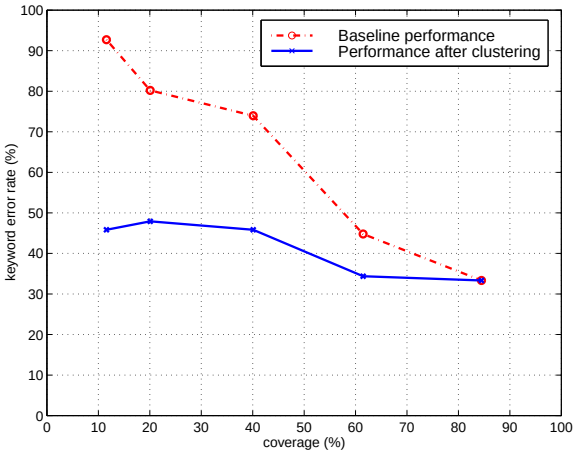

Figure 9-3: Keyword error rate of baseline recognizer and clustering recognizer as total coverage varies.

$$KER = \frac{F + M}{T} * 100 \tag{9.1}$$

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 9-3 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.

"destroy"	"green"	"landmine"	"robot"	"spaghetti"	"yellow"
[d ih s t r ao]	[g r iy n]	[l ae d m ay n]	[r ow b ao]	[s p ix g eh t iy]	[y eh l aw]
d ih s t r oy	g r iy n	n ae n s m ay n	r ow b ao n	t ax g eh t iy	y eh n l ow
d ih s t r ay	g r iy n	l ae d m ay n	r ow b ao	s p ix g eh t iy	y ae l ow
s t r ao	g r iy d	l ae n m ay n	r ow b aw	s p iy t ax	y eh l ow
dh ax s t r ao	r iy n	l ae n m ay n	m ow b ao	d ix g eh	y ax l aw
dh ax s t r oy	d r iy n	l ae d m ay n	r ow v ae	d ix g ih	y eh l aw
d ih s t r ao	g r iy	m ae d m ay n	r aw b ao	s p ix g eh d t iy	
d ey s t r ao d	g r iy	l ae d s m ay n	r ow b aa		
dh ey s t r ao	g r iy n	l ae d m ay n	r ow b aa		
d ih s t r ao	g r iy d	l ah n n ay	r ow b ah		
d ih s t r ay	g r iy d	l ae n t w ay n	r ow w ae		
	k r iy n	n ae n d ix n l ay n			
	r iy	b l ae n t w ay n			

Figure 9-4: Converging on a vocabulary. The top row shows English words that were spoken to the robot repeatedly. The second row shows the phonemic version of those words the robot chose. The remaining rows show the transcripts of each individual utterance. The version chosen by the robot is what it *speaks*, however it will recognize words close to any of the variants as corresponding to the same word. So if the person says "spaghetti" and the robot hears [d ix g ih], then it will recognize and mimic that word as [s p ix g eh t iy]. Clearly this will limit the size of the robot's vocabulary, but that seems a necessary trade-off with the current state of the art.

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.

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 for the work described in (Breazeal and Aryananda, 2000). Semantically salient words such as "kismet", "no", "sorry", "robot", "okay" appeared among the top ten clusters.

9.5 Real-time vocabulary extension

In actual operation, ideally new vocabulary items could be added instantaneously, rather than extracted through a slow offline procedure. To achieve this, the robot was given a much simpler vocabulary extension mechanism, where novel isolated words are mimicked back immediately by the robot and added to its lexicon. As words are heard, they are grouped based on a weak measure of similarity – the statistics of pairs of phonemes two words have in common (See Figure 9-4). Similar-sounding words will be merged, but this can be accepted since it would be difficult to reliably differentiate them anyway. In a sense, the robot will only permit sufficiently different sounds to converge to a vocabulary. This method is appropriate if the desired working vocabulary at any point has a relatively small number of words. This is all that can really be supported in a noisy environment without user-specific training or well-placed microphones, anyway. Initially this behavior was achieved by using the dynamic vocabulary API of IBM ViaVoice. It proved simpler to use raw phonemic recognition and an external Viterbi alignment procedure, although ideally this would be merged with the speech recognition system for optimal performance.

9.6 Stabilized perceptual interface

Just like the object recognition system discussed in Chapter 5, recognized words were communicated to the rest of the system using a stabilized interface. As vocabulary items are created, they are assigned a unique 'feature line.' The meaning of that feature line is conserved as much as possible from then on, so that the line will respond to the same situations in future as it did in the past. Offline clustering is done to refine the online vocabulary grouping. This is initially done without any regard to the stabilized interface so that off-the-shelf clustering algorithms can be used; then as a last step models are compared with previous models and aligned appropriately.