How can a dialogue system compensate for speech recognition deficiencies?

Rebecca Jonson
Department of Linguistics, Göteborgs University
rj@ling.gu.se

January, 2002

Abstract

Speech recognition performance affects the impression of a dialogue system as a whole and for that reason it is important to find methods that can compensate for recognition deficiencies. Spoken language recognition is an extremely difficult task, but to make a dialogue system perform well does not mean recognising exactly what the user has said but to be able to extract the correct semantic content from the input. In this paper, I will give an overview of different ways of enhancing speech recognition in the context of dialogue systems.

1 Introduction

McTear [McTear, 2001] includes in his list of issues of importance in future spoken dialogue research, the issue of investigating how language understanding and dialogue management can compensate for deficiencies in speech recognition.

Although speech recognition has made important improvements during the last decade, recognition performance for dialogue systems is still often quite poor. Recognition failures affect the performance of the whole dialogue system and therefore many researchers have tried to find methods that can improve upon the speech recogniser output. I agree with McTear that this challenging task needs more research, and in this paper I will try to sketch out different strategies that could be used.

Before starting to explore the different compensation methods that are being used in the research area and could be used in the future, we will start to look at which factors in spoken language that influence the recognition performance.

2 What makes speech recognition for dialogue systems so difficult?

Spoken language involves several factors that affect the speech signal and make it difficult for machines to recognise speech. Analysis of continuous speech is per se a difficult task as there are no clear boundaries between words, and the simplified definition we make of phonemes is much more difficult due to phenomena such as coarticulation.

The variation of human voices is immense and we have to try to cope with all kinds of speakers that utter the same thing in acoustically quite distinct ways. Even the same speaker varies acoustically depending on physical and emotional factors such as having a cold or being stressed. Individual variations are not only on the voice level but speakers also behave different in dialogues and have different talking habits e.g. more or less disfluent speech production.

In addition, a recogniser has to be able to distinguish speech from other acoustic signals such as noise. If it was the case that the acoustic environment would be held constant this could be modelled, but systems are exposed to many different environments and channels which are hard to handle acoustically. In [Blomberg and Elenius, 1997] channel bandwidth, noise, room acoustic and telephone and microphone frequency are mentioned as some of the disturbing factors on the speech signal.

Apart from these factors, people do not speak as clearly and eloquently as they think they do, but produce filled pauses, repetitions, repairs, utter truncated words, make false starts, make mistakes and slips of the tongue and even change their minds during speech production. How shall we cope with all these disfluencies? We also produce a lot of extralinguistic sounds such as inhalations, smacks or coughs. We have seen that it is difficult to model non-verbal disturbing sounds. Imagine the difficulty to make a machine distinguish between linguistic and extra-linguistic verbal signals! A small comfort is that it seems that users shape up a little bit in dialogues with machines and speak clearer than they should have done with a human dialogue counterpart. The occurrence of disfluencies is still common though in human-machine dialogues as the Adapt Corpus shows [Gustafson *et al*, 2000] but are less frequent than in human-human dialogues as the disfluency study made by Eklund and Shriberg [Eklund and Shriberg, 1998] shows.

Another aspect is the problem of vocabulary. People will always come up with words that developers had not thought of i.e. out of vocabulary words (OOV). How shall we handle unknown words? One could of course propagate for larger vocabularies, but a larger vocabulary also complicates the recognition task as the more words the more probability that there are words acoustically similar to confuse the input word with. In addition, the lack of good methods to identify OOV, which instead leads to recognition of an in-vocabulary word, also affects the recognition of surrounding words.

As Gold and Morgan [Gold and Morgan, 2000] point out humans can often compensate for all of these disturbing factors on the speech signal whereas today's speech recognisers cannot. Humans in contrast to automatic recognisers easily distinguish speech from noise, can recognise unknown words and understand incorrect utterances with help of contextual interpretations. Lippman [Lippman, 1997] shows in his study how humans outperform machines in recognition tasks and how humans do not seem to be affected by noise in great extent. He also points out that further studies need to be made to clarify how humans proceed to compensate for all disturbance on the speech signal. There seem to be a lot of things going on in human perception and one thing is sure, we need to integrate speech recognition closer with other parts of a dialogue system to be able to do just as humans do: use other sources, such as context knowledge, in the disambiguation task of the speech signal.

3 How to compensate for speech recognition deficiencies when building a dialogue system

In the following sections, I will discuss and exemplify how we with different methods and techniques may compensate for some of the disturbing factors mentioned above and by that improve upon the output from a speech recogniser.

3.1 Input modalities

Dictation systems in contrast to recognisers for dialogue systems, are sold with a microphone and usually include a training procedure for the user. Speech recognisers for dialogue systems get their input in several ways depending on the application, but most commercial systems take their input via the telephone line. By that, they have to take into account the distortion of the signal in the telephone line and be prepared to receive signals both from terrestrial network and the mobile network. Instead of trying to cope with all kinds of input modalities we could try to impose a single input modality on future users. Consider that we would use our mobile, outfitted with good microphones which we know how to use and are aware of closeness to microphone etc. Factors such as noise would of course still be affecting the signal but we would at least have a single channel, and users would always use the same input modality and hopefully learn how to use that input modality in a successful way.

3.2 Acoustic models and language models

In contrast to dictation systems, where the user train the system to his voice, speech recognition for dialogue systems is often speaker independent and does not even take into account that it is the same user during the whole dialogue. In some domains it would be possible to have a short training procedure e.g. in systems that already have speaker verification. In some applications it is actually possible and necessary to know the identity of the user by telephone number recognition or speaker verification and in this case we should not neglect the techniques used in dictation systems but we could make use of user-adapted acoustic models.

Lippman's study [Lippman, 1997] shows how humans adapt their perception to the talker, channel and talker-style variability using only short speech segments. So what we would need for recognisers are dynamically adaptive acoustic and language models. We have seen that speaker variation is one of the factors that complicates the recognition task so a desirable strategy is to take into account that it is the same speaker during the whole interaction, and adapt during dialogue to the user and not consider each utterance as an utterance from a new speaker. This is an issue that Young [Young, 1996] brings up and he points out that this could reduce error rates considerably, especially for atypical speakers.

Studies of humans' language modelling capabilities have also been done and compared to the performance of trigram language models [Lippman, 1997]. Even here humans outclass today's methods, and we need to produce language models with much lower perplexities or combine language models with other methods. Jelinek [Jelinek, 1991] pointed out that after decades of progress in speech recognition, trigram models were still much the same and although the weaknesses of the trigram models are known, improvements on them had come up short.

At least for languages with restricted word order, language models capture, if trained on a large corpus, a lot of information, i.e. syntactic, semantic and pragmatic information. One way to get down perplexity is to combine word based language models with so called class based models ([Jurafsky and Martin, 2000]), that can either be POS-classed trained on a POS-tagged text or the classes may be semantic e.g. all proper names have the same probability.

The disadvantage of language models is that they do not capture long-distance dependencies. The use of one content word often gives raise to the use of another. In human speech perception we use what Jurafsky et al. [Jurafsky and Martin, 2000] call semantic word association to access words that are related to preceding words more quickly. A method used to get away from this lack of long-distance semantic dependencies in language models is to combine n-gram models with language modelling based on what is called *latent semantic analysis* [Zang and Rudnicky, 2002]. By using this statistical technique it is possible to capture correlation of content words [Zang and Rudnicky, 2002].

Speech recognition for dialogue systems has during the latest years been focused on grammar-based approaches although language models seem to have a better overall performance [Gorrell *et al*, 2002]. This probably depends on the time-consuming work of collecting corpora for training the language models compared with the more rapid development of grammars. However, language models are more robust, can handle out-of-coverage output, perform better in difficult conditions and seem to work better for naive users [Knight *et al*, 2001]. On the other hand, the grammar-based approach in the experiments reported in [Knight *et al*, 2001] outperforms the language model approach on semantic error rate on in-coverage data. As seen, the two approaches both seem to have their advantages. An interesting approach, as proposed by Gorrell et al. [Gorrell *et al*, 2002], would therefore be to combine grammars and language models, and e.g. rely on a more robust statistical language model when the grammar fails. Other researchers try to improve upon language modelling by incorporating syntactic structure to complement the locality of trigram models [Jelinek and Chelba, 1999]. In [Jelinek and Chelba, 1999] a language model that uses grammatical analysis to predict the next word, a so called statistical Structured Language Model, is described which has led to decreasing perplexity and word error rate. Hybrid approaches seem to be the future.

Another issue is to what extent we should adapt language models to the users. Such models would need to reestimate words that are used more frequently. We could also think of the possibility of adding new words, e.g. by spelling. How do we estimate the probability for these words? If we were only allowed to add words of certain semantic classes the word would get the probability of its class in a semantic class based model.

Unknown words is a huge problem for recognisers that humans seem to have little problem with. Out of vocabulary words (OOV) will always appear even if a large corpora have been used and developers have struggled hard to predict user behaviour. The most important thing is to model OOV correctly so that they don't disturb the rest of the input i.e. so that the system can recognise at least the known words correctly. The system ability to catch up new words seems to me a very hard problem although we would like to handle this at least with clarification questions where the user is asked to spell the unknown word. Purver [Purver, 2002] describes how such a process of unknown words could proceed in a dialogue system.

3.3 N-best hypotheses

Let us move on from the speech recognition module and instead try to process the recognition results that we have got. Many recognisers do not give a single transcription of what has been said but provide what is called an N-best sentence hypotheses, i.e. a list with several alternative transcriptions of what may have been uttered. However, the recogniser's top choice is often not the correct transcription but hypotheses that have been rated lower by the recogniser are many times more correct. In the corpus used in [Quesada *et al*, 2002] 12% of the times the correct recognised utterance was included in the N-best list but not as the top-choice. This shows that if we could identify these correct alternatives in the N-best list we would be able to make a significant improvement in recognition performance. The number of hypotheses produced by recognisers vary but tests have shown that it is only worth working on the top-choices due to the complexity of the task [Chotimongkol and Rudnicky, 1998].

More or less advanced post-process methods have been used to analyse and decide on the best choice from the N-best list. The simplest way is to take the top choice, parse it and in case the output is satisfactory skip the rest. In case the top choice fails to give a useful parse then the next sentence in the list is parsed and so on, until a set number of alternatives have been gone through. Although this is a very simple strategy it actually improves upon concept accuracy. More advanced methods take into account several of the top-choices and decide with different analysis methods on the most prominent choice. Improvement on word error rate of about 35% was made in Communicator [Chotimongkol and Rudnicky, 1998] by reordering the top-5 hypotheses from a 25-best list with a linear regression model considering information of domain grammar and previous system utterances. In the SIRIDUS project [Quesada *et al*, 2002], context is taken into account to choose between alternatives by using the information held in the Information State (IS). For example, in their House Simulator, the system might prefer "switch on" to "switch off" when knowing that the device in question is currently off [Quesada *et al*, 2002]. Parsing techniques have also been used to choose between the word sequences generated by recognisers giving higher score to grammatically correct hypotheses [van Noord *et al.*, 1997].

Instead of working with N-best rescoring, some researchers have chosen to work with what is called word lattice rescoring [Quesada *et al*, 2002], [Jelinek and Chelba, 1999]. A word lattice is, as described in [Gold and Morgan, 2000], a graph of possible word sequences, with associated probabilities from the on-line acoustic and language models. Some recognisers provide such lattices and make it possible to, instead of waiting for the recogniser to generate an N-best list, compute on the lattice and extract paths from it. In this way it is possible to find candidates that is not part of the N-best list and it is possible to extract those using linguistic knowledge in an early process step. In the Siridus Project the first experiments have shown that this strategy, using linguistic parsing in the recognition procedure, may improve upon recognition performance.

3.4 Dynamic language models and grammars

In contrast to dictation systems speech recognisers for dialogue systems actually have access to contextual information, i.e. knowledge of the situation and the dialogue in process, that could be used to recover or reduce recognition errors. This advantage has not been commonly explored. In a dialogue system we have information about the state of the dialogue, the preceding system utterances which of course should not be neglected in the recognition process.

In dialogue systems with a directed dialogue, moving from state to state, it is possible to use different language models for each state that apply to the current situation. Changing language models on the fly will improve recognition accuracy as language models can be held state-specific. For example if we are asking for a telephone number we can use a language model specifically developed for that recognition task. This method can be used even in less directed dialogue, e.g. mixed-initiative dialogues by defining states as preceding system prompts. Dialogue state adaptation of the language models was applied on the CMU communicator system and gave a word error rate reduction of more than 10% [Xu and Rudnicky, 2000].

In the same way we can change between different semantic grammars to interpret an utterance considering only that specific state context. The combinatory strategy of grammars and language models mentioned earlier could also be used this way, by having a state specific grammar for recognition with a general language model backing up in case the grammar fails and to allow the user to move outside the grammar's bounds. Another way to use the knowledge of preceding system prompts is to regenerate language models and change probabilities depending on the state. For example if the system has asked for a departure city then the semantic class "city" in a class based language model should be reestimated with a higher probability.

3.5 Natural language understanding

Suppose we have received the best possible output from the recogniser by using some of the techniques mentioned above including choosing the best of alternatives in the N-best list or by finding a good path in the lattice. The work is far from done, even if we have managed to get the word error rate lower we need to use a good extraction procedure of semantic concepts to obtain high concept rate and by that succeed with the dialogue.

Dialogue systems use widely differing methods to interpret the users' utterances ranging from key-word spotting to advanced parsing techniques. Keyword spotting may actually be sufficient for some applications but if we aim for more advanced human-language interaction we need more sophisticated methods. Existing parsing techniques have primarily been developed for parsing text and do not seem to be suitable for spoken language as they expect features of the input that spoken language does not live up to, such as grammatically correct input. The output from a recogniser is many times fragmentary and ungrammatical, due to the nature of spoken language and because parts fall off in the recognition process.

Recent research has focused on more robust parsing techniques which accept the characteristics of spoken language. Semantic grammars are commonly used instead of traditional syntactic grammars in dialogue system interpretation modules. The rules and constituents in a semantic grammar as described in [Jurafsky and Martin, 2000] correspond to entities and relations from the domain in discussion.

Another approach is to extract the dialogue act or dialogue move from the output to retreive the user's intended dialogue action. Fukada et al. [Fukada *et al*, 1998] describe a probabilistic method for dialogue act extraction, using a speech act dependent n-gram model. This proposed method performs better and is more robust when compared to a grammar-based approach.

Hybrid approaches have also been suggested with deep parsing combined with shallow parsing, or more robust techniques, to back off to when deep parsing fails [Wang *et al*, 2002], [van Noord *et al.*, 1997]. Wang et al. [Wang *et al*, 2002] combine statistical techniques with a semantic context free grammar in a two stage understanding module which have led to a significant performance improvement. The statistical task classifier is applied first to obtain the task class, as it has been shown that this method had lower task classification rate as it is more robust. After that, the grammar is applied to do a more deep parsing only using the grammatical rules related to the identified task class.

Whatever method chosen for interpretation we do need more research on how interpretation and recognition can be further integrated. One thing is clear: spoken language definitely needs more study and we have to get away from the text-dependency in this research area. Perhaps spoken language is not ungrammatical but it is just that we have not figured out the syntax of spoken language including disfluency patterns?

3.6 Natural language generation

It may seem far-fetched that the system's output can improve upon recognition but as Glass [Glass, 1999] points out *the precise wording of the response can have a large impact on the user response.* Studies have shown that people tend to build up a shared terminology during interaction. This phenomena is called *lexical entailment* [Gustafson *et al*, 1997]. So, how does the system's output affect the user's responses? Well, it seems to be the case that the introduction of a word in the system's response raises the probability of the use of that specific word in the user response. Studies of the Waxholm corpus at KTH [Gustafson *et al*, 1997] showed that people imitated and caught up the vocabulary used by the system. It is therefore important to carefully choose the words to use in the system utterances, and assure that the system can recognise these words. A word used by the system could also be assigned a higher probability than a possible synonym. We could thereby rescore the probability of the words we use.

The strategy in Natural Language Generation is then to make the user say what the system knows it can handle. Vague system responses give larger variation of input whereas more directed responses decrease the user response variety. Another issue is when recognition starts to fail, in that case the system could reformulate its utterances using a different vocabulary, and see if the user adopts the new vocabulary instead of the one that seemed to fail for that specific person. For example the user may pronounce a specific word in an atypical way but the user's pronunciation of a similar expression is easy to recognise.

A lot of research effort is put on error recovery strategy and error detection.

Supposing the dialogue system is capable of knowing if a conversation is on track or what things are going wrong then we could indicate to the user how to speak when things go wrong. If the user starts talking louder, to clear or to use out of vocabulary words etc. In that case the system could try to make the user aware of this. In [Gorrell *et al*, 2002] such a strategy is used by giving relevant help messages that can put a user back on track and increase the possibility that subsequent utterances will be recognised correctly. When understanding fails the system indicates this to the user, but instead of just saying “please rephrase” or similar, the system gives intelligent help and makes suggestions of how the user should proceed dependent of what the system thinks the user tried to do.

Another aspect, apart from improving upon recognition, is to improve upon a dialogue system’s knowledge of how well the recognition process is doing and then adapt different dialogue strategies depending on how confident the system is that the recognised utterance was what the user uttered or how confident the system is on each word in the hypothesis. Different methods on how to estimate confidence scores have been proposed (see e.g. [Carpenter *et al*, 2001]) but the most interesting part is how to use these confidence scores. Larsson [Larsson, 2002] proposes three levels of grounding strategies: optimistic, cautious and pessimistic grounding depending on the reliability of the recognition. E.g. if the confidence score is low the system could adapt a more cautious grounding strategy by asking the user to confirm explicitly what has been understood. This does not improve on recognition per se, but would improve the robustness of the dialogue and perhaps the user’s impression of the recognition process. It could also lead to that the recognition of subsequent utterances would improve.

As mentioned earlier, speakers do not only have distinct voices but do also behave differently in dialogue and have different levels of experience with the system. User-adaptive systems could take advantage of the information they have about the current user and its behaviour and adapt the dialogue strategy and help messages to the user which could lead to improvement of the recognition for that specific user.

4 Discussion

The fact that speech recognition is an extremely difficult task is probably not a surprise to the reader. In this paper, I have tried to gather some of the methods that are used in dialogue system research to compensate and recover from bad speech recognition performance. As seen, the methods are numerous and apply to different levels of a dialogue system, ranging from adapting acoustical models to choosing different speech generation strategies.

Many methods and strategies are still in the cradle and need to mature to be used in commercial systems but there seem to be many ways to improve upon speech recognition and there is not one single solution that should be applied but combinations of different methods will be needed.

Although word error rate will not decrease radically, concept error rates will definitely be reduced. Is not that what we aim for in dialogue system development, to achieve task success and more satisfied users?

References

[Blomberg and Elenius, 1997] Mats Blomberg och Kjell Elenius (1997) *Automatisk igenkänning av tal*

[Carpenter *et al*, 2001] Carpenter, Paul., Jin, Chun., Wilson, Daniel., Zhang, Rong., Bohus, Dan., Rudnicky, Alex. *Is this conversation on track?* Proceedings of Eurospeech 2001 (Aalborg, Denmark), pages 2121-2124.

[Chotimongkol and Rudnicky, 1998] Chotimongkol, A. and Rudnicky, A.I. (1998). *N-best Speech Hypotheses Reordering Using Linear Regression* Proceedings of Eurospeech 2001 (Aalborg, Denmark), pages 1829-1832.

[Eklund and Shriberg, 1998] Eklund, R. and Shriberg, E. (1998). *Crosslinguistic Disfluency Modeling: A comparative Analysis of Swedish and American English Human–Human and Human–Machine Dialogs* Proc. Intl. Conf. on Spoken Language Processing, vol. 6, pp. 2631-2634, Sydney, Australia.

[Fukada *et al*, 1998] Fukada, T., Koll, D., Waibel, A. and Tanigaki, K. (1998). *Probabilistic Dialogue Act Extraction for concept based multilingual translation systems* Proceedings of the ICSLP 98, December, 1998.

[Glass, 1999] J. Glass *Challenges for Spoken Dialogue Systems* Proc. 1999 IEEE ASRU Workshop, Keystone, CO, December 1999.

[Gold and Morgan, 2000] Ben Gold and Nelson Morgan, 2000. *Speech and Audio Signal Processing: Processing and Perception of Speech and Music* John Wiley and Sons, Inc.

[Gorrell *et al*, 2002] Gorrell, G, Lewin, I and Rayner, M. 2002. *Adding Intelligent Help to Mixed Initiative Spoken Dialogue Systems* Proceedings of ICSLP 2002.

[Gustafson *et al*, 2000] Gustafson, J, Bell, L, Beskow, J, Boye, J, Carlson, R, Edlund, J, Granström, B, House, D and Wirén M (2000) *AdApt - a multimodal onversational dialogue system in an apartment domain* In Proc of ICSLP 2000, Beijing, 2:134-137

[Gustafson *et al*, 1997] Gustafson, Larsson, A., Carlson, R., Hellman, K. (1997) *How do System Questions Influence Lexical Choices in User Answers?* Eurospeech '97.

[Jelinek, 1991] Jelinek, Frederick. *UP FROM TRIGRAMS! The struggle for improved language models* in Proceedings EUROSPEECH 91, pp. 1037-1040, Genova, (Italy), 1991.

[Jelinek and Chelba, 1999] Jelinek, Frederick, Chelba Ciprian. (1999). *Putting language into language modeling.* in Proceedings of EUROSPEECH 99.

[Jelinek and Chelba, 1999] Jelinek, Frederick, Chelba Ciprian. (1999). *Recognition Performance of a Structured Language Model.* in Proceedings of EUROSPEECH 99.

[Jurafsky and Martin, 2000] Jurafsky, D., Martin, J. H. (2000) *Speech and Language Processing* Prentice Hall

[Knight *et al*, 2001] Knight, S, Gorrell, G, Rayner, M, Milward, D, Koeling, R and Lewin, I. 2001. *Comparing Grammar-Based and Robust Approaches to Speech Understanding: A Case Study* Proceedings of Eurospeech 2001

[Larsson, 2002] Larsson, Staffan. *Issue-based Dialogue Management.* PhD Thesis, Göteborg University. 2002

[Lippman, 1997] R.P. Lippman *Speech recognition by machines and humans* Speech Communication vol 22 no 1, pp 1-15. 1997

[McTear, 2001] Michael F McTear (2001) *Spoken dialogue technology: enabling the conversational influence.* Submitted to ACM Computing Surveys.

[Purver, 2002] Matthew Purver. *Processing Unknown Words in a Dialogue System* In Proceedings of the 3rd ACL SIGdial Workshop on Discourse and Dialogue, pages 174-183, Association for Computational Linguistics, July 2002.

[Quesada *et al*, 2002] Quesada J.F., Amores, J.G., Manchón, P., Peréz, Knight, S., Milward, D., Thomas, J. (2002). *Possibilities for Enhancing Speech Recognition by Consulting Information States* Deliverable D2.3 Siridus project.

[van Noord *et al.*, 1997] van Noord Gertjan, Bouma Gosse, Koeling Rob, Nederhof, Mark-Jan. *Robust Grammatical Analysis for Spoken Dialogue Systems* Journal of Natural Language Engineering, 5(1), 1999, pages 45–93 (written 1997).

[Wang *et al*, 2002] Wang, Ye-Yi., Acero, Alex., Chelba, Ciprian., Frey, Brendan., Wong, Leo. *Combination of Statistical and Rule-based approaches for Spoken Language Understanding* in Proc. Int. Conf. on Spoken Language Processing. Denver, Colorado, Sep, 2002.

[Xu and Rudnicky, 2000] Xu, W. and Rudnicky, A. (2000) *Language modeling for dialog system?* Proceedings of ICSLP 2000 (Beijing, China). Paper B1-06

[Young, 1996] Young, SJ (1996). *Large Vocabulary Continuous Speech Recognition* IEEE Signal Processing Magazine 13(5): 45-57.

[Zang and Rudnicky, 2002] Zang, Rong. and Rudnicky, Alexander I. (2002) *Improve Latent Semantic Analysis based Language Model by Integrating Multiple Level Knowledge* Proceedings of ICSLP 2002 (Denver, Colorado), pages 893-896.