Asian Journal of Information Technology 12 (1): 7-13, 2013
ISSN: 1682-3915
© Medwell Journals, 2013

Analysis of Various Efforts to Compensate for Automatic Speech Recognition Deficiencies in Spoken Dialogue System

[1]S. Lokesh and [2]G. Balakrishnan
[1]Department of CSE, Anna University, Chennai, Tamil Nadu, India
[2]Indra Ganesan College of Engineering, Tiruchirappalli, Tamil Nadu, India

Abstract: One of the draw back in Spoken Dialogue System through speech is the brittleness of Automatic Speech Recognition (ASR). ASR Systems often unpredicted the user input and they are unreliable when it comes to judging, recognition failures and lacking in estimating the own performance of an interaction system. Humans overtake ASR Systems on most tasks related to speech understanding. One of the reasons is that humans make use of much more knowledge. For example humans appear to take a variety of knowledge-based aspects of the current dialogue into account when processing speech. The main purpose of this study is to investigate whether speech recognition also can benefit from the use of higher level knowledge sources and dialogue context when used in human computer interaction. This review provides more insight into what type of knowledge sources in spoken dialogue systems would be potential contributors to the task of ASR and how such knowledge can be represented computationally. The purpose of this survey was also to exemplify the difficulties that arise when using speech in dialogue systems. Many of these difficulties have also been encountered in the experiments.

Key words: Automatic Speech Recognition (ASR), dialogue system, Human Computer Interaction (HCI), knowledge-based, India

INTRODUCTION

Speech recognition performance has gradually improved over the past 20 years or so. The main reason is probably more work out and more computing storage for more training data. Improvements of the ASR technology have also been made with improved feature extraction techniques, new noise robustness techniques, improved acoustic modeling and improved language modeling. Humans outperform ASR on all types of recognition tasks. Most difference in performance is found when it comes to spoken dialogue and noisy environments. On the Switchboard corpus (Godfrey and Holliman, 1997) which is a spoken dialogue corpus, Lippmann (1997) reports a 43% Word Error Rate (WER) for ASR but only 4% WER for human subjects. Today, the best ASR error rate for the Switchboard corpus is probably a bit lower but still orders of magnitude higher than the reported human error rate above. This comparison is not truly fair as the vocabulary size and the knowledge sources used are not comparable for humans and machines. Humans can make use of higher level knowledge such as meanings of words, situational context or dialogue context. Even when these sources are not involved, however such as in the recognition of non-sense words and non-sense sentences, human performance is unrivalled (Lippmann, 1997). This indicates that humans use other acoustic cues than the ones that are used in ASR. Humans can use all the information of the acoustic signal while ASR is restricted to the features researchers extract in the DSP phase. Therefore, there seem to be some important cues missing even in the feature extraction step. The superiority of human performance in comparison to ASR indicates that there is much room for improvement on different levels.

In the meantime, researchers attempt to compensate for the deficiencies in ASR by incorporating knowledge about human spoken processing and language into the current statistical framework. The following sections consist of a survey of such ideas, techniques and approaches that have been used by researchers on different levels in the ASR framework in an attempt to enhance ASR. The survey will also focus on the problems arising when using ASR in dialogue systems.

IMPROVING THE FRONT END

Dictation systems in contrast to speech recognizers for dialogue systems are sold with a microphone usually include some guidance of how to use the system and also

Corresponding Author: S. Lokesh, Department of CSE, Anna University, Chennai, Tamil Nadu, India

require a training procedure with the user. Guiding users on how to give speech input to a computer can save much misrecognition. A training procedure to adapt to a speaker's voice leads to important recognition performance improvement. In most speech solutions for customer service systems where any customer should be able to call, it is hard to include any training procedure and adaptation as many of the callers will not use the system again. Also, such a procedure would require some sort of speaker identification for successive calls to be able to reuse adapted models. In addition, as customers expect an immediate service they are probably reluctant to lose time training the system or listening to instructions on how to best give input to ASR.

Speech recognizers for dialogue systems get their input in various ways, depending on the application. Most commercial systems get their input via a telephone line. In these systems, the speech recognizer must take into account the distortion of the signal in the telephone line, the limited bandwidth and be prepared to receive signals both from the terrestrial network, the mobile network and more recently also IP telephony. This limits the recognition performance in contrast to dictation systems with a direct input line from a headset.

ROBUSTNESS TO NOISE

ASR is very sensitive to adverse environments and performance degrades considerably. Humans on the other hand, easily adapt to unexpected conditions. Researchers identify two kinds of noise: stationary and non-stationary. Stationary noise such as a computer fan or a car engine is easier to model than non-stationary noise such as a door slam or a mobile ringing. The attempts to make ASR less sensitive to noise, so called noise robustness techniques, constitute an intensive research area. Hung *et al*. (2001) gives a good overview of noise robustness techniques in ASR with primary focus on robustness to stationary noise. Noise robustness techniques include everything from coping with noise by developing better microphones, handling echo-cancelling, finding noise resistant signal features in DSP, subtracting noise from the speech signal in the DSP to adapting acoustic models to different noise conditions.

IMPROVING DIGITAL SIGNAL PROCESSING

On the digital signal processing level, knowledge about human auditory perception has improved ASR considerably and is currently used to greater or lesser extents in different ASR Systems. The assumption is that as speech is intended for human hearing which is limited, researchers should not consider what humans do not hear but focus on the perceived parts. Unfortunately, researchers do not have a deep insight into human auditory processing. As perception is an internal human process it is hard to study and the knowledge that researchers have, has been drawn from experiments with human subjects.

INSPIRATION FROM HUMAN AUDITORY PERCEPTION

The most successful way of using knowledge about human auditory perception in ASR has been by using techniques that are inspired by the non-linear human perception of frequency bands. In cepstral analysis, features are transferred to a Mel scale which corresponds better with the human perception of frequencies. Mel Frequency Cepstral Coefficients (MFCC) is now a days the most common feature representation in ASR. Perceptual Linear Prediction (PLP) is a feature extraction technique with a more direct relation to human hearing with non-linear frequency scale, equal loudness curve (Hermansky, 1998). This has been shown to give a system more robustness to noise and channel distortions as well as speaker differences (e.g., adult vs. child speech). In this way, PLP has been a way to obtain acoustical features that are more resistant to variation.

AUDITORY PERCEPTION AND ARTICULATORY PRODUCTION

From experiments on human subjects, it has been shown that human auditory perception and human articulatory production seem to be somehow intertwined. For example, Wilson *et al.* (2004) showed how the articulatory part of the brain is activated when listening to speech. Another example comes from listening experiments with speech synthesis where subjects feel a sensation of exhaustion when TTS Systems speak faster than humans are able to do due to breathing constraints. Human articulatory production seems to focus on phonetic contrasts rather than absolute phonetic targets. Perception seems to have been adapted to the limitations of the speech apparatus and to focus on contrasts. There seems to be a trade off between production and perception where speakers try to minimize the energy consumption while maintaining the spoken signal perceptual distinctive on demands of the listener. This leads to the broad variation of acoustic realizations of speech sounds.

IMPROVING ACOUSTIC MODELLING

In contrast to dictation systems where the user trains the system to her voice, ASR for dialogue systems is often speaker independent and does not even take into account that it is the same speaker during the whole dialogue. The study by Lippmann (1997) shows how humans adapt their perception to the speaker, channel and speaking style using only short speech segments. So, what researchers would want for recognizers are somehow dynamically adaptive acoustic models.

ADAPTING TO THE SPEAKER

Researchers have seen that speaker variation is one of the factors that complicate the recognition task. Recognition performance differs tremendously between different speakers. For some speakers, ASR just does not work properly whereas for others it works reasonably well. It is common to talk about speakers as sheep and goats where sheep are the good ones and goats the ones who perform badly (Doddington *et al*., 1998). A desirable strategy for SDSs would be to take into account that it is the same speaker during the whole interaction and be able to adapt to the user. However, speech recognizers usually consider each utterance as an utterance from a new speaker. This is an issue that Steve and Chien (1996) brings up and he points out that this could reduce error rates considerably, especially for atypical speakers (i.e., goats). In some applications, it is actually possible and necessary to know the identity of the user by telephone number recognition or speaker verification. In this case, researchers should not neglect the enrollment techniques, used in dictation systems but make use of User-Adapted Acoustic Models. The two baseline systems in this review could well be used by one single person on their laptop and it would therefore be possible to train the speech recognizer on their voice. This would most certainly lead to improved recognition performance.

As reported in Lippmann (1997), humans seem to adapt to a new speaker after hearing only three syllables. The current state of the art in automatic adaptation techniques, still needs as much as 10 sec of speech. What these techniques adjust are the GMMs by adapting the mean values in the GMMs to the speaker's voice. Unfortunately, the time this adaptation currently takes is too long for many commercial dialogue systems where perhaps the total time for the user turns are expected to be around 10 sec of speech.

SPEAKING STYLE

Researchers have already mentioned that the recognition performance degrades heavily in SDSs in real situations. As discussed, one of the reasons is background noise and disturbance to the acoustic signal. However, if researchers take the Let's Go Public System as example again, it was reported that even when taking away utterances with crosstalk and background noise the WER was still 60% as compared to 17% in laboratory settings (Raux *et al*., 2005). This indicates that there must be other factors involved. In an experimental study presented by Weintraub *et al*. (1996), recognition of spontaneous dialogue was compared to read speech. The channel, the speakers and the words spoken were held invariant by having subjects interacting spontaneously and then ask them in a subsequent experiment to read the transcriptions of their own spontaneous utterances. The difference in recognition performance for these two tasks was huge with a 53% WER for the spontaneous utterances whereas only 29% when read. Even, when subjects were asked in a third task to read in a conversational style the error rate was much lower (38%) than for the spontaneous task. This indicates that there must be something about the way of speaking in dialogue that complicates the recognition task. Speakers vary in speaking style depending on the task, the situation and the acoustic environment. Lindblom (1990) defines speaking style as going from hyper speech to hypo speech. Spontaneous dialogue would be more on the range towards hypo speech whereas read speech would be closer to hyper speech. Hypo speech is characterized by a high speaking rate, less careful speech that leads to more reductions and more coarticulation. This makes the pronunciations of words more diverse. It has been shown that especially highly predictable words vary in their pronunciations as there is no real need for the speaker to articulate these well. In the switchboard corpus, it was found that the pronunciation of words was extremely varied. There were for instance 100 different ways of pronouncing the word. One approach would be to add pronunciation variants to the HMM lexicon. In most ASR Systems the developer can add and modify pronunciations of words. However, adding more pronunciations may lead to more ambiguity and a deterioration of the search process (Soltau and Waibel, 2000).

IMPROVING LANGUAGE MODELLING

Jelinek (1991) pointed out years ago that after decades of progress in speech recognition SLMs or specifically trigrams were still much the same. Although, the weaknesses of the trigram models were known, improvements on them had come up short. Jelinek was not alone in proposing the search for more sophisticated language modelling techniques. Brill *et al*. (1998), Moore (1999) and Glass (1999) all proposed the use of more linguistic knowledge in SLMs. Attempts at alternative ways of modelling language other than with SLMs have been scarce. One of the few alternatives uses Artificial Neural Nets (ANN) to build language models (Xu and Rudnicky, 2000). Xu and Rudnicky (2000)'s experiments show that ANNs can learn to model language with comparable performance to SLMs but with a much higher computational cost. A more successful but limited alternative in SDSs have been the use of SRGs for smaller tasks and in cases where training data has not been accessible. Today, almost two decades after the publication of up from Trigrams (Jelinek, 1991), researchers are still mostly using trigrams. According to a survey among speech scientists by Moore (2005) SLMs are expected to persist.

Statistical language modelling and human language modelling actually seem to have some things in common. SLMs are built on word frequencies just as human lexical access seems to be based on frequency. Similarly to SLMs, humans seem to make predictions of coming words based on previous words. In speech, humans even shorten more predictable words (reductions) while putting longer duration on infrequent words. Both humans and SLMs also seem to process multiple words in parallel (Jurafsky and Martin, 2008). However, statistical language modelling suffers from several problems:

- Data sparseness
- Restriction to very local word context
- Difficulty of adding or detecting new vocabulary
- Difficulty of adapting to different language contexts
- Static frequencies which do not rely on the bigger context

DATA SPARSENESS

SLMs suffer from data sparseness when there is not enough appropriate data to be able to estimate good probabilities of words and word co-occurrences. This is very common in SDSs where in-domain data is seldom available and spoken corpora are rare. When data is sparse an SLM will obtain low estimates for many word occurrences and will most probably not have been exposed during training to many of the words that it will encounter when used. The most commonly used smoothing techniques (or discounting algorithms) are Good-Turing, Witten-Bell and Kneser-Ney (Stolcke, 2002). In addition, techniques for combining higher and lower ordered n-grams are used such as Katz-Backoff and deleted interpolation. These are applied to be able to rely on lower order n-grams (e.g., bigrams) when a higher ordered n-gram (e.g., a trigram) is not encountered in the model to be able to estimate the probability of the higher ordered n-gram. The difference between these last two techniques is that deleted interpolation also rely on lower ordered n-grams for non-zero counts whereas Katz-Backoff only use the information from lower ordered n-grams for zero counts (Hung *et al*., 2001).

LONG DISTANCES

At least for languages with more restricted word order, trigram SLMs seem to capture, if trained on a large corpus both syntactic, semantic and pragmatic information (Jelinek, 1991). However, language is much more complex than three-word sequences. In human speech perception, researchers exploit relations between the meanings of words in order to be able to prime future occurrences of words in a given context. One statistical technique to be able to capture correlations between content words is Latent Semantic Analysis (LSA), a.k.a. Latent Semantic Indexing (Zhang and Rudnicky, 2002). This technique is widely used in the information retrieval community in an attempt to structure the relationships among words by reducing dimensionality. A matrix of word co-occurrences is built up. To reduce the dimensions of such a matrix an algorithm called Singular Value Decomposition (SVD) is used (Bellegarda, 1998). The use of LSA in ASR in combination with n-grams have led to reduction both in perplexity and WER when compared to n-gram models on the WSJ corpus. Genevieve (2006) shows how LSA does not depend on SVD but can be used with a different algorithm: Generalized Hebbian Algorithm (GHA) (Gorrell and Webb, 2005). SVD and GHA were used in Gorrell (2007) to show the value of decomposition in statistical language modelling. It was shown to be hard to obtain a good performance with language models with reduced dimensionality alone. However, when interpolating them with standard n-grams an important reduction in perplexity could be shown. For large domains there is a tractability issue as these models are computationally expensive to produce. However, as Gorrell (2007) points out for smaller domains such as in SDSs this approach could be of interest although it has not yet been examined.

NEW VOCABULARY

The vocabulary in SDSs will probably never fully cover a user's needs. Unknown words or so called OOVs usually appear even if a large corpus has been used and developers have struggled hard to predict user vocabulary. In fact, users are very creative or rather language is rich which means that earlier unseen words will most probably appear. The number of unseen words in a test set is measured as the OOV rate. The OOV rate affects the recognition performance significantly. With a bigger vocabulary researchers have more chance of covering more of the user's vocabulary. However, the size of the vocabulary also affects recognition performance. The bigger the vocabulary the bigger the search space and the more room for ambiguity and failure as there will be more words acoustically similar to confuse the input word with. Also if there are more words in the vocabulary a bigger corpus will be needed to get good estimates of all these words in different contexts. A vocabulary which is too large may slow down the recognition process and actually lead to more errors.

Unknown words are hard to tackle for recognizers whereas humans seem to have little problem detecting them and are often also able to recognize them. Although, the automatic recognition of novel words is desirable the most critical point in ASR is to detect OOVs correctly as they lead to misrecognitions. When users make use of words unknown to the ASR System it will try to match these two words in its predefined vocabulary. As language models are built on the probability of word occurrences such an incorrect recognition may therefore also affect the recognition of surrounding words.

DEVELOPING LANGUAGE MODELS FOR NEW DOMAINS

SLMs are unfortunately very bound to the training data and very sensitive to new types of data. It is therefore hard to reuse SLMs from one domain to another or adapt them to a new purpose. The mismatch can either be in style or in content. A mismatch in speaking style could be for example using newspaper text to build a model for a broadcast news recognition task. A mismatch in content could be to use transcriptions from spoken interactions in a travel domain for a tax office domain. Dialogue system developers are often confronted by the dilemma of a small amount of in-domain corpus material and large amounts of other corpus material. However, somehow there should be something generic, domain-independent in all the amount of text we have that we could reuse. As an example some phrases such as I want to seem to be quite common in many spoken dialogue system domains. Researchers have therefore attempted to create language models based on a mix of topics that are expected to model what is generic in a language and does not vary from one application to another (Solsona *et al*., 2002). The idea is that such models can then be adapted to different domains and tasks by combining them with domain data.

In dialogue systems with a directed dialogue where users are guided from state to state, it is possible to use different language models in each state. An approach in many commercial systems to constrain ASR and thereby improve the accuracy is to use state specific models. Such models will only be able to recognize a restricted set of utterances and words specific for the current dialogue state.

IMPROVING ASR HYPOTHESES SELECTION

A straightforward approach to testing new techniques or additional knowledge sources in ASR has been to apply them in a post-process step on the output from the speech recognizer, e.g., on the N-Best. In this way, there is no need to integrate proposed techniques for example more Sophisticated Language Models, into the internal recognition process to be able to evaluate them. Techniques are evaluated by their success in selecting the best possible hypothesis from N-Best lists when re-ranking (also reordering or rescoring) the hypotheses in N-Best lists. The meaning of the best possible hypothesis can either be the hypothesis that would minimize the WER (best word sequence match) or the hypothesis that best captures the user's intention (minimizing the CER). The recognizer's top choice is sometimes not the most accurate option but hypotheses that have been rated lower by the recognizer can be more accurate. In the corpus used by Quesada *et al*. (2002), it was estimated that 12% of the time the correct recognition of the utterance was included in the N-Best list but not as the top ranked item. In communicator corpus, a human-machine spoken dialogue corpus, a 37% relative improvement in WER would be possible if an Oracle Method existed to pick the best hypothesis from 25 best lists. For the switchboard corpus a 26% relative improvement in WER is reported as the upper bound (the Oracle rate) (Brill *et al*., 1998). A 59% relative improvement in WER was reported as possible on 10 best lists from the ATIS corpus using a bigram model in (Manny *et al*., 1994). These figures indicate that if researchers could identify the correct alternatives in N-Best lists we would be able to make a significant improvement in recognition performance.

To investigate the limits and possibilities of improving recognition with the use of N-Best lists researchers have given humans the task of re-ranking the outcome of speech recognizers (Brill *et al*., 1998). In human subjects were given the task of selecting hypotheses that they thought would have the lowest WER from 10 best lists for three different speech recognition tasks (Switchboard, Broadcast News and Wall

Street Journal). The purpose of the study was to explore what linguistic knowledge humans make use of when carrying out such a task as well as to estimate the possible gain. The subjects were also allowed to edit the hypotheses. For each N-Best list they were asked to determine what knowledge or information they had used for their decision. Human subjects were indeed able to improve on the output of all three recognizers. Taking into account the possibility of editing the improvement was even better. The most complicated task was shown to be the spoken dialogue task, switchboard where the gain was lower. This was probably because the higher error rate of the recognizer for this task which did not leave enough cues to work on in the hypotheses (Brill *et al*., 1998). According to the subjects the most common knowledge/information that they had used (for the spoken dialogue task) was the choice of words in closed classes (e.g., that vs. than) and open classes and the completeness of the sentence. For the broadcast news and wall street journal tasks the choice of determiners and prepositions had an important influence. Apart from linguistic knowledge the subjects also stated that they had made use of world knowledge in their selections.

IMPROVING CONFIDENCE ANNOTATION

Confidence scores measure the reliability of the correctness of recognition results. The output from ASR Systems is undoubtedly uncertain and error-prone. ASR Systems output the most likely word sequence among its possible word sequences but do not tell us how well that word sequence matches what the user actually said. Confidence scoring concerns estimating the extent to which the words in a hypothesis match what was actually said by giving a score of reliability to each word in a hypothesis. Such word confidence scores are also often used to estimate an utterance score to reflect the reliability of the whole hypothesis (the utterance). As reflected in the earlier study improvement on confidence scoring have sometimes been related to N-Best hypothesis selection as re-estimated confidence scores have been used to rescore (or re-rank) the lists. In this way, better confidence measures can also lead to better hypotheses selection. If confidence scoring is not reliable for example high scores are given to misrecognized utterances (or words), a SDS will be incapable of dealing with both correct and incorrect utterances. Knowledge of the reliability of a hypothesis is crucial in dialogue systems to be able to properly decide what to do with a hypothesis. The most evident decision-making in SDSs is the binary decision of accepting correctly recognized hypotheses and rejecting wrongly recognized hypotheses. In dialogue systems, researchers want to avoid the rejection of correct recognitions, False Rejections (Frs) as well as avoid the acceptance of misrecognitions (False Acceptances (FAs)). The most common approach when using ASR confidence scores is to set a threshold and accept hypotheses with a confidence score above that threshold and reject hypotheses below it.

CONCLUSION

The purpose of this survey was also to exemplify the difficulties that arise when using speech in dialogue systems. Many of these difficulties have also been encountered in the experiments. Although, the dialogue system interactions recorded with the systems have not been carried out in a real environment they were conducted outside the laboratory. Subjects used a headset and a laptop and included both experienced and inexperienced dialogue system users. All of them were informed about how to use the headset to not speak too low or to mumble too much. This was to avoid worst case scenarios. However, as the reader will see in the reports of the experimental data, recordings include noise, crosstalk and disfluencies and are thus far from clean speech. This means researchers also had to cope with these problems in some way.

REFERENCES

Bellegarda, J.R., 1998. Multi-Span statistical language modeling for large vocabulary speech recognition. Proceedings of the International Conference on Spoken Language Processing, December 4, 1998, Sydney, Australia, pp: 2395-2399.

Brill, E., R. Florian, J.C. Henderson and L. Mangu, 1998. Beyond N-grams: Can linguistic sophistication improve language modeling? Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics-Volume 1, August 10-14, 1998, Morgan Kaufmann Publishers, San Francisco, CA., pp: 186-190.

Doddington, G., W. Liggett, A. Martin, M. Przybocki and D. Reynolds, 1998. Sheep, goats, lambs and wolves: A statistical analysis of speaker performance in the NIST 1998 speaker recognition evaluation. Proceedings of the 5th International Conference on Spoken Language Processing, November 30-December 4, 1998, Sydney, Australia.

Genevieve, G., 2006. Generalized hebbian algorithm for incremental singular value decomposition in natural language processing. Proceedings of the 11st Conference of the European Chapter of the Association for Computational Linguistics, April 3-7, 2006, Trento, Italy, pp: 97-104.

Glass, J., 1999. Challenges for spoken dialogue systems. Proceedings of IEEE ASRU Workshop, December 1999, Keystone, CO., pp. 307-310.

Godfrey and E. Holliman, 1997. Switchboard-1 release 2. Linguistic Data Consortium, Philadelphia.

Gorrell, G. and B. Webb, 2005. Generalized hebbian algorithm for latent semantic analysis. Proceedings of the 9th European Conference on Speech Communication and Technology, September 4-8, 2005, Lisbon, Portugal, pp: 1325-1328.

Gorrell, G., 2007. Generalized hebbian algorithm for dimensionality reduction in natural language processing. P.hD. Thesis, Linkoping University, Sweden.

Hermansky, H., 1998. Should recognizers have ears? Speech Commun., 25: 3-27.

Hung, X., A. Acero and H.W. Hon, 2001. Spoken Language Processing, a Guide to Theory, Algorithm and System Development. 1st Edn., Prentice Hall Inc., USA., ISBN-10: 0130226165, pp: 980.

Jelinek, F., 1991. The struggle for improved language models. Proceedings of Eurospeech, September 24-26, 1991, Genova, Italy, pp: 1037-1040.

Jurafsky, D. and J.H. Martin, 2008. Speech and Language Processing. Prentice Hall, New Jersey.

Lindblom, B., 1990. Explaining Phonetic Variation: A Sketch of the H and H Theory. In: Speech Production and Speech Modeling, Hardcastle, W.J. and A. Marchal (Eds.). Kluwer Academic Publisher, The Netherlands, pp: 403-439.

Lippmann, R.P., 1997. Speech recognition by machines and humans. Speech Commun., 22: 1-15.

Manny, R., D. Carter, V. Digalakis and P. Price, 1994. Combining knowledge sources to reorder N-best speech hypothesis lists. Proceedings of the Human Language Technology Workshop, September 1, 1994, Plainsboro, NJ., pp: 219-221.

Moore, R.C., 1999. Using natural-language knowledge sources in speech recognition, in computational models of speech pattern processing. http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.27.2954.

Moore, R.K., 2005. Spoken language processing: Piecing together puzzle. Speech Commun., 49: 418-435.

Quesada, J.F., J.G. Amores, P. Manchon, K. Perez, S. Milward and D. Thomas, 2002. Possibilities for enhancing speech recognition by Consulting Information States. Deliverable D2.3, SIRIDUS, http://citeseer.uark.edu:8080/citeseerx/showciting;jsessionid=4038B13447D6A5C09C0CCD2ED1FA9B90?cid=684160.

Raux, A., B. Langner, D. Bohus, A.W. Black and M. Eskenazi, 2005. Let's go public! Taking a spoken dialog system to the real world. Proceedings of the Interspeech'2005-Eurospeech, 9th European Conference on Speech Communication and Technology, September 4-8, 2005, Lisbon, Portugal, pp: 885-888.

Solsona, R.A., E.F. Lussier, H.K.J. Kuo, A. Potamianos and I. Zitouni, 2002. Adaptive language models for spoken dialogue systems. Proceedings of the International Conference on Acoustic Speech and Signal Processing, Vol. 1, April 2007, Orlando, Florida, pp: 37-40.

Soltau, H. and A. Waibel, 2000. Specialized acoustic models for hyperarticulated speech. Proceedings of the International Conference on Acoustics, Speech and Signal Processing, (ASSP'00), Istanbul, Turkey, pp: 1779-1782.

Steve, G. and J.T. Chien, 2012. Large-vocabulary continuous speech recognition systems: A look at some recent advances. IEEE Signal Proces. Magazine, 6: 18-33.

Stolcke, A., 2002. SRILM: An extensible language modeling toolkit. Proceedings of the International Conference on Spoken Language Processing, September 2002, Denver, CO., pp: 901-904.

Weintraub, M., K. Taussig, K. Hunicke-Smith and A. Snodgrass, 1996. Effect of speaking style on LVCSR performance. Proceedings of International Conference on Spoken Language Processing, October 3-6, 1996, Philadelphia, PA., pp: 16-19.

Wilson, S.M., A.P. Saygin, M.I. Sereno and M. Iacoboni, 2004. Listening to speech activates motor areas involved in speech production. Nat. NeuroSci., 7: 701-702.

Xu, W. and A. Rudnicky, 2000. Can artificial neural networks learn language models. Proceedings of the International Conference on Spoken Language Processing Vol. 1, October 13-15, 2000, Beijing, China, pp: 202-205.

Zhang, R. and A.I. Rudnicky, 2002. Improve latent semantic analysis based language model by integrating multiple level knowledge. Proceedings of the International Conference on Spoken Language Processing, September 2002, Denver, CO., pp: 893-896.