Chapter 1

CHALLENGES IN SERVING SPEECH SOLUTION FOR USE BY MASSES

Arun Kumar Pande, Sunil Kumar Kopparapu*
TCS Innovation Labs - Mumbai,
Yantra Park, Tata Consultancy Services, Thane (West), Maharastra, INDIA.

PACS Keywords: Speech Technology, Speech for Masses, Speech Recognition.

Key Words: **A**MS Subject Classification:

Abstract

Speech technology has never been in more demand than before with the percolation of mobile phones into the bottom of the pyramid (BoP). In a multi-lingual, multi-accent country like India, speech becomes the only *natural* communication channel for the masses to transact, though there are constraints in terms of availability of corpus, noisy telephone lines, socio-economic issues etc. To establish a working speech solution which can work for the BoP requires innovative and novel use of the state of art speech recognition research and development. In this Chapter, we first establish the research challenges that exist in serving the masses and show how these challenges can be effectively overcome using an example.

*E-mail address: {Arun.Pande,SunilKumar.Kopparapu}@TCS.Com

1. Introduction

Speech technology has never been in more demand than before with the percolation of mobile phones into the bottom of the pyramid (BoP). According to a recent report [1] at the end of 2010 there was an estimated 5.3 billion mobile cellular subscriptions worldwide and as much as 90% of the world population and 80% of the population living in rural areas has access to mobile network. This coupled with the exponential growth in primarily consumer-oriented services industries has fueled the need for consumers to access information. The outcome of these two makes speech based solutions the need of the day. The mobile phone growth providing the channel and the services industry providing the need.

Speech technology has a rich history[2] and has been continuously been researched in literature and only of-late is it being deployed outside laboratories in practice and commercial applications. The increase in computing power has significantly helped speech research and has provided the necessary fuel to be able to recognize spoken words and phrases with good accuracies. However, the ultimate aim of being able to recognize *naturally spoken speech* nurtures research in the area of speech technology and continues to hold interest of the research community worldwide. While the aim to create an AI based naturally spoken speech recognition remains in the forefront; to address commercial viability and actual deployment in the field effective approaches have been adopted [3]. Menu based speech solution systems which need to understand only a closed set of words or phrases (often called limited vocabulary) have been adopted instead of having to try and understand a naturally spoken speech. Though this has the disadvantage of not being user friendly [3] it has the benefit of building and deploying commercially viable and usable speech solution. This is a classic case of divide and rule, when technology is not mature enough to address the requirement, you put constraints of the requirement and use the available technology. Along these lines are to build a speech solution to only address a specific domain or an application so that the number of words or phrases that the speech engine has to recognize are greatly reduced thus transforming a difficult problem into a problem which is less difficult. While natural language speech interfaces are the talk in the research community, yet in practice, menu based speech solutions thrive. Typically in a menu based speech solution the user is required to respond by speaking from a closed set of words when prompted by the system. A sequence of human speech response to the IVR prompts results in the completion of a transaction. A transaction is deemed successful if the speech solution can correctly recognize all the spoken utterances of the user whenever prompted by the system.

The menu based approach, enables use of the current state of the art speech technology and this in turn has made it possible to develop and deploy speech solutions in commercial world in the western countries with actual demonstration of return on investments [4]. Most of these deployments have been for transactional purposes for financial institutions [5] and travel [6] and have been crafted to take care of the deficiencies of the speech recognition by language modeling (LM) and restricting the vocabulary. With increasing use of mobile phones the voice based transactions has become even more important and necessary because of the voice channel availability on person 24×7.

While it is believed that the speech technology can work well for India demographically the speech recognition history in India has been limited to research in laboratories and a couple of research laboratories. And unfortunately attempts to deploy speech recognition

based solutions commercially is both sparse and when deployed is of limited use. From the Indian scenario speech based solution is very apt because in India,

1. majority of this population are in need of some transactional information (travel inquiry, news, stock quotes, yellow pages etc) or the other very often.
2. a large percentage of the population is not English literate;
3. no major access to broadband or Internet or computers;
4. they can speak their native language but might not be able to write or read in their native language;
5. people have access to mobile phone thanks to the proliferation of mobile phones in the last couple of years;
6. sending SMS in local Indian language is still cumbersome even to the people who can write in their language;

In total sum, in the Indian context, there exists a very huge opportunity for speech based solutions to thrive especially if they can address usage by masses. However, there are India specific challenges. The fact that there are so many languages and dialects spoken and hence to be addressed by the speech solution adds to the fact that a majority of the population is English illiterate makes it neither easy nor technologically feasible to replicate even the most *successful* speech solutions that have proven in Western countries [7], [8], [9], [10]. Speech solutions for Indian scenario need India specific modifications to address India specific idiosyncrasies. Irrespective of the state of the art of the technology and the numerous challenges that speech recognition faces in an Indian scenario, the need for speech recognition is both urgent and widely required. One of the prime drivers for the need of speech is that the penetration of computers and landlines in India has been poor but on the other hand the penetration of mobile phones has changed the communication landscape significantly. In that sense a voice channel as a mode to communicate with the IT systems is open for exploitation.

From the economics perspective, the Gross domestic product (GDP) has been very impressive in the last couple of years in India but this has been restricted to the urban areas and hence a small percentage (30 %) of the total population. The majority of the population, namely, 70% is rural and its contribution to GDP is 22% only. The GDP urban-rural imbalance has left low growth in rural areas. One of the fall out of this is the social tensions. The main cause of this can be attributed significantly to the lack of a channel for information flow from places of information mostly in urban areas to the locations where this information is urgently required. By and large the only *natural* channel is the speech and speech recognition technology is an answer to many inefficiency issues in rural sector of Indian economy. The rest of the Chapter is organized as follows: We dwell on the challenges of speech recognition in the Indian scenario in Section 2. In Section 3. we suggest the approach that needs to be adopted to enable building speech based solutions for the masses. In Section 4. we describe the applications that are suitable for mass usage and give some example and finally conclude in Section 5.

2. Challenges of Indian Speech Recognition

The mobile phone proliferation has had no rural-urban divide and hence the ownership and use of mobile phone as a device to communicate is pan India. While the literacy level is low in the rural block, the confidence in using the latest technology is high [11, 12] and as a consequence the mobile technology has been adopted with ease. But what actually divides urban-rural categories is the literacy in English both spoken and written and largely combined with the lack of literacy in written native language. This literacy situation makes the use of any channel other than voice a sore point in a nicely laid mobile network landscape. For example any transaction requires the origin of a query from the user; *use of voice* (may be English in the urban area and a native Indian language in the rural area) for querying is probably the only channel which cements the rural-urban divide. Even if transaction was possible on the mobile phone on the SMS channel, the urban-rural divide still exists because unlike English entering non-English text is both laborious and time consuming [13] in any of the Indian languages.

2.1. Cultural

There is a huge cultural aspect that comes into play as far as adaption of a technology is concerned how ever useful the technology might be. In the Indian scenario, the cultural, literacy and economical conditions of the masses have a strong resistance to digest using the latest technology, such as speech based solutions. Culturally, it is not acceptable to be speaking to a machine for information or for conducting transaction. This cultural perception is changing with the introduction of mobile phones. The mobile phone as a portable personal device stays with the person almost all the time has eased the challenge of acceptance of mobile technology. However there are challenges in terms of designing speech solution interfaces that find acceptance for mass usage. Dialog design and persona of a speech solution can play a very crucial role in acceptance of the speech solution by the masses. Traditionally, IVR which required the person to use the keypad of the telephone did not find acceptance in small towns and villages because culturally people are used to *ask* questions and get them answered and not so much as be shown a path to fetch their answers by pressing a couple of keys. Additionally ofcouse pressing keys on a mobile phone is difficult because every time you have to press a key you need to move the mobile phone from the ear to a position where one can see the keypad.

2.2. Language and People Diversity

India officially recognizes 22 languages and the fact of the matter is that there are hundreds of dialects and thousands of accents. To cater to these diversity would be a huge task even if one restricted to one language that is widely used, say Hindi[14]. To address one language would mean one would still have to address the several dialects and accents. However, to truly address the masses, the challenge of 22 official languages is to be addressed by quickly adapting a speech recognition solution working in one language into another second language and deploying it in the second language. While this has been addressed in [15] it still happens to be a detrimental factor in launching a speech solution in several different languages. because of large grammar variations within Indian languages.

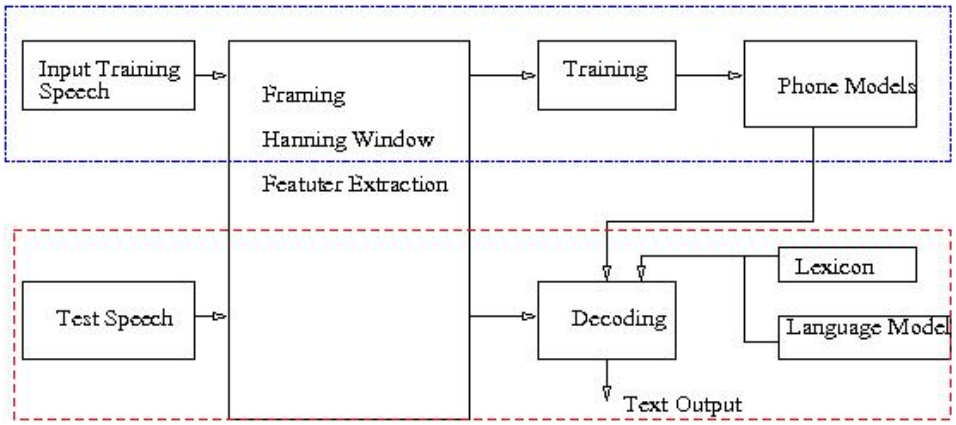

Figure 1. Speech Recognition Overview

With urbanization and geography shift of people the ability to converse in many languages is becoming very common. and this in recent times leads to an issue of mixed language. A very large population of people are using mixed language in everyday conversation without actually being *aware* of its usage. Use of mixed language is found to be very common in the young [16]. Mixed language arises through the fusion of two or more, usually distinct, source languages, normally in situations of bilingualism, so that it is not possible to classify the resulting language as belonging to either of the language families that were its source [17, 18, 19]. Though mixed language is defined as a mixture of two distinct languages void of the knowledge of which language is mixed into which, at least in the Indian context, the native language is the primary language and the non-native language is the mixed or the secondary language.

In addition to the language issues, scaling up a speech solution to address a diverse population is a challenge in itself especially in terms of voice biometrics which is an essential ingredient of a speech solution to enable transaction based queries of the user.

2.3. Other

Infrastructural challenges in the form of noisy environment and noisy telephone lines, especially the quality of landlines in the rural area. While the mobile phone voice quality is good the environment from which this device can be used makes the speech quality poor for any speech recognition to happen reliably. Even in urban areas there is a lot of babble noise because of the crowded public places making speech from that area prone to bad speech recognitions.

For a typical speech recognition process (see Fig. 1), acoustic models are built using speech corpus during the training phase (boxed blue -.-. in Fig. 1). In the testing (boxed red - - - in Fig. 1) or the recognition phase, these acoustic models are used along with a pronunciation lexicon and a language model to recognize speech. The language model and the pronunciation lexicon also require the text corpus corresponding to the speech or the domain and needs to be built. So typically, a speech corpus is a must be build a speech recognition platform. Lack of a standard speech corpus for the Indian languages and the

challenges associated with building a speech corpus for so many languages is a major challenge.

3. Indian Speech Recognition - Approaches

Speech based solutions are probably most needed in India as has been seen in the previous sections but this pressing solution need is packed with challenges that need to be addressed. Probably the best way to approach is to first tackle one language, say, Indian English. There is a vast majority of India in urban area that speaks English and this would mean that it would have a fewer rural specific challenges. Develop one speech based application which can demonstrate ease of usability combined with commercial benefit. This commercial success would get many private companies that are not currently too sure if they need to invest or not in speech in India to go ahead and make some significant investment to get speech recognition to work in the Indian context. In addition to an in-depth study into speech signal processing in the Indian context in terms of the study of phonemes and also the specific language models availability of a speech corpus is a must to enable build speech solutions.

3.1. Building Speech Corpus

Availability of a speech corpus for a specific language is an essential requirement to build a speech recognition engine and all speech recognition based solutions thereof in that language. The process of creating a speech corpus in any language is a laborious, expensive and time consuming process. We propose a method to create speech corpus which is frugal in all the three senses, namely, it is less expensive, less laborious and less time consuming. Typical speech corpus is a set of speech files and the associated transcriptions. The usual process of speech corpus creation starts with the linguist determining the language specific idiosyncrasies. Then a textual corpus is built to make sure that the phonemes in the language are uniformly distributed (also called phonetically balanced corpus). Then the target speaker age, accent, gender distribution is computes and then the actual speech recording is done from the chosen speakers in predetermined environments. Typically, the text corpus is created keeping in mind the underlying domain for which the speech recognition is going to be used. Once the speech corpus is collected, the speech is transcribed manually. The whole collection of the speech data and the corresponding transcription together forms the speech corpus. This elaborate process means several languages do not have a speech corpus available specifically when that language does not have commercial speech recognition based solution viability. The next section tries to give a method that enables creation of a speech corpus in a frugal way, thus making it possible to construct speech corpus.

3.1.1. Possible Approach

Speech data is available on the web, especially in the form of news (example, [20]) which is accompanied by the transcripts (example, [21]). In a way one has access to a well transcribed speech corpus however there are certain limitation in terms of (a) limited speaker variability (number of speakers), (b) limited environment (recording environment) and (c)

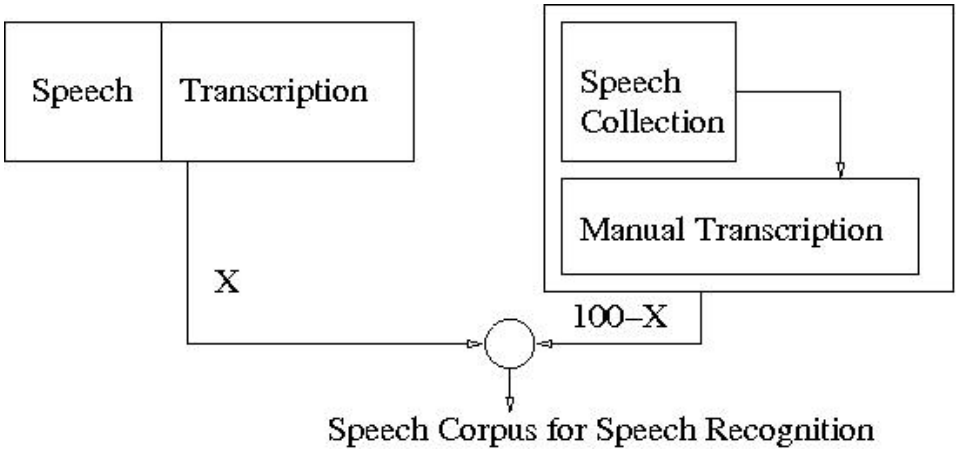

Figure 2. An Approach for frugal speech corpus creation

limited domain (because you choose from what you have). The idea is to bank on this already available speech corpus and to try and create the phonetic balance of the speech corpus by collecting *minimal* speech data keeping in mind the need for variability in terms of environment, gender, age and dialect. The combination of available speech data and a smaller amount of collected data enables construction of a speech corpus for a given language in a frugal way. Let us say we need to create a speech corpus for a language L. We identify sources on the web which have public access to speech data. An automatic process can download the speech data (example [20]) and the corresponding transcription (example [21]). We use speech segmentation algorithm to match the transcription to the speech file at a sentence or word level. Now we analyze the transcripts using language processing tools [22] to identify those text segments that would satisfy the phonetic balancing of the speech corpus. A large portion, say X % of the speech corresponding to the collected text can come from the speech data already available and the remaining $(100 - X)$ % could be collected in the usual way. The choice of X would determine the amount of effort, time and expenditure in constructing the speech corpus. The larger the X the more frugal the construction of the speech corpus. In the limiting case, when $X = 0$, it would be what is conventionally used for speech corpus creation. On the other extreme, $X = 100$ one would have access to the cheapest mode of creation of speech corpus at the cost of lack of diversity in terms of speaker variability, environment. One could control X based on where one would like to use the speech corpus for speech recognition engine training. Figure 2 depicts the suggested approach. The left had side shows the use of already available resources on the web and the right hand side shows the conventional speech data collection process. The fact X determines the amount of deviation from the conventional approach in terms of making it frugal.

3.2. Speech Recognition Platform

Given the complexity of the issues surrounding the speech recognition based solutions for the Indian scenario it would be but natural to expect that no one speech recognition solution

ASR-1	How much	the fund shall	I get in case I	can sit in	My booked	to dictate
ASR-2	How much	Refund shut	I get in case I	canceled	my book	ticket
Fused	How much	the fund shall	I get in case I	canceled	my booked	ticket

Figure 3. Fusing two ASR outputs to produce better speech recognition accuracies

might work for different situations. Figure 3 shows the example of the output of two speech recognition engines. The first row is the recognition due to a speech recognition engine (say ASR-1) and the second row is the output of the speech recognition engine (ASR-2). Clearly the two recognitions for the same spoken query */How much fund shall I get in case I canceled my booked ticket/*[1], are wrong. While ASR-1 recognizes it as "How much the fund shall I get in case I can sit in My booked to dictate", ASR-2 recognizes it as "How much Refund shut I get in case I canceled my book ticket". Individually, both of the outputs are in error, however by fusing the outputs of ASR-1 and ASR-2, out could get a reasonably accurate recognition of the spoken query. Figure 4 shows a framework that could be used which could perform *better* than any one ASR. As seen in Figure 4 one of the

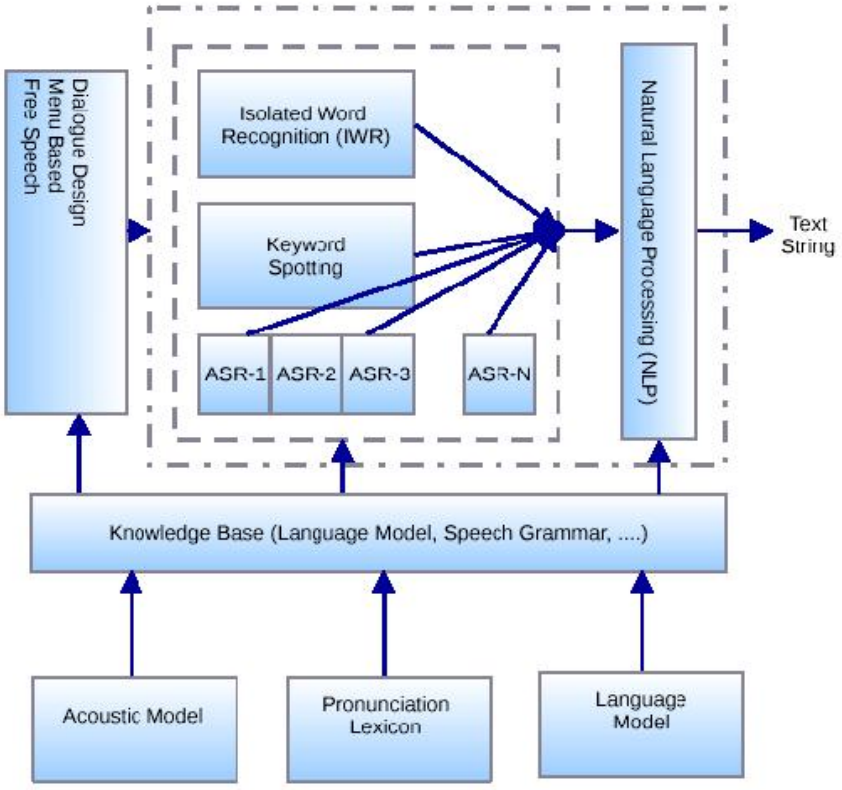

Figure 4. Robust Speech Recognition

[1]// indicates the spoken sentence, namely, /S_1/ is the spoken equivalent of the written sentence S_1.

ways to achieve a working level speech recognition accuracies would be to integrate several speech recognition engines (say ASR-1, ASR-2 etc) in addition to using an isolated word recognition (IWR) engine and key word spotting (KWS) engines. These engines would be amply supported by different acoustics models (created from speech corpus) and language models (created from transcripts). A domain based natural language processing (NLP) would assist in fusing the output of the several speech recognition engines and in addition would perform the task of intelligently extracting the intent or meaning associated with the speech input.

Given the complexity of the problem in hand, it might be necessary not to visualize *one* solution to all the problems as is done in the case of a *product approach* of building a speech recognition engine; rather in case of the Indian scenario one would expect a *solution based* approach where each problem is addressed separately and as a whole. In a very general situation one could have a very domain specific speech recognition platform rather than a single speech platform that works for all the domains.

4. Applications for the Masses

Majority of India lives in the rural area and the majority of the information lies in the urban area and there is no real way to bridge the gap unless a pipe is created in the form of a channel that allows the rural folks to get information that exists but elsewhere. Essentially most applications that cater to the masses are of the information seeking type and restricted to a very few areas like agriculture (for example, */Where can I sell my agri product so that I can get a good price?/ /Which is the best pesticide to use?/*, */Do you think it will rain in the next 48 hours?/*, */What is the price of potato in the market closest to me?"/*), general information (example, *How late is this train?*, */Is there a government policy that enables me to get a loan at concessional rates?"/*, */Is there accommodation available on this train on this date?/*). In the absence of this information the rural folks are powerless in their decision making. A speech based application which can cater to bridging the gap between the information sink and the information source can change a whole lot in terms of rural GDP and living standards. The other area that could have a large impact is the rural banking. In India, today a large population is unbanked, this is not so much because of the smaller economy in the rural area, in fact 20% GDP contribution comes from rural economy which could translate to 200 billion dollar economic activity. The main reason banks have not been able to tap into this huge market is because of the commercial viability of setting up banks in rural areas. Specifically speech recognition and speaker verification can play important role and allow banks to cater to towns and villages without actually having to open branches in every town and village.

4.1. Example

We describe how a speech based solution can be deployed to assist establishing a channel for information disseminate to the rural folks. Indian Railway [23] is one of the most used modes of transportation in India with 9000 passenger trains running each day covering 64015 kilometers and moving over 20 million passengers. Information about trains is one of the sought information especially in peak seasons. The need for information has resulted

in the Indian railway setting up a manned call center. Setting up a call center to handle these calls (8 lakh calls per day) is an expensive proposition and speech based solution is seen as the *natural* alternative.

While the magnitude of the problem to be addressed is huge especially in terms of being able to offer this information about trains to the masses, it can be broken down into addressable solutions. For example, though there are 9000 trains that the speech recognition engine might have to recognize, this number can be significantly reduced if you consider the fact that not all trains pass through all the railway stations and the fact that people geographically *close* to a certain railway station do not need information about the trains that do not pass through that station. This observation enables one to build a reliable and usable speech solution to address the problem of enabling masses to ask for information. Further, one could use a language specific speech recognition depending on the region in which the speech solution is deployed. For example, in a particular geography there could only be a small set of languages (much less than the official number of languages 22) that would be spoken, making speech engine region (and hence language) specific instead of building speech recognition engine to recognize all languages simultaneously. Using this philosophy, we are in the process of building a speech recognition system that can be used by masses to get information about Indian railway.

5. Conclusions

One is on the look out for a platform which allows a person to pick up a phone and get the desired information by asking for the information in a natural way. While the technology that can make this possible is still developing, there are a number of innovative and workable approaches that can be adopted to make speech the channel for transaction and query answering. In this Chapter, we have shown through examples how this can be made possible after identifying the problems associated with building working solution for use by masses. There are several direct and indirect impact due to use of speech as a channel to enable information reach. In a agricultural scenario, information about commodity prices in real time gives power of negotiation to the rural farmer for getting better prices for their produce.

Acknowledgments

The authors would like to thank members of the TCS Innovation Labs - Mumbai for the numerous simulating interactions that have directly or indirectly found mention in this Chapter.

References

[1] ITU, “The world in 2010: ICT facts and figures.” [Online]. Available: http://www.itu.int/ITU-D/ict/material/FactsFigures2010.pdf

[2] B.H.Juang and L.R.Rabiner, *Automatic Speech Recognition - A Brief History of the Technology*, 2005. [Online]. Available: www.ece.ucsb.edu/Faculty/Rabiner/ece259/Reprints/354_LALI-ASRHistory-final-10-8.pdf

[3] S. K. Kopparapu, "Voice based self help system: User experience vs accuracy," in *SCSS (1)*, T. M. Sobh, Ed. Springer, 2008, pp. 101–105. [Online]. Available: http://www.springerlink.com/content/j8772401lr126qp7/

[4] D. Fluss, "The intimate connection between customer satisfaction and ROI." [Online]. Available: http://www.cisco.com/en/US/prod/collateral/voicesw/custcosw/ps5694/ps1006/prod_white_paper0900aecd800e9d7a.pdf

[5] Salmat, "Suncorp - success story." [Online]. Available: http://www.vecommerce.com/assets/successstories/SuncorpCaseStudy.pdf

[6] Voxeo, "Case study : Prophecy IVR hosting." [Online]. Available: http://www.voxeo.com/pdf/1800FlightsCaseStudy.pdf

[7] Y. Sun, J. Gemmeke, B. Cranen, L. Bosch, and L. Boves, "Using a dbn to integrate sparse classification and gmm-based asr," *Proceedings of Interspeech 2010*, 2010.

[8] X. Lua, S. Matsudaa, M. Unokib, and S. Nakamuraa, "Temporal contrast normalization and edge-preserved smoothing of temporal modulation structures of speech for robust speech recognition," *Speech Communication*, vol. 52, pp. 1–11, 2010.

[9] Y. Zhao and B. Juang, "A comparative study of noise estimation algorithms for vts-based robust speech recognition," *Proceedings of Interspeech 2010*, 2010.

[10] C. Kim and R. Stern, "Feature extraction for robust speech recognition based on maximizing the sharpness of the power distribution and on power flooring," *IEEE International Conference on Acoustics Speech and Signal Processing*, pp. 4574–4577, 2010.

[11] S. Lobo, P. Doke, and S. Kimbahune, "Gappagoshti: a social networking platform for information dissemination in the rural world," in *Proceedings of the 6th Nordic Conference on Human-Computer Interaction: Extending Boundaries*, ser. NordiCHI '10. New York, NY, USA: ACM, 2010, pp. 727–730. [Online]. Available: http://doi.acm.org/10.1145/1868914.1869015

[12] E. Bellman, "Rural india snaps up mobile phones." [Online]. Available: http://online.wsj.com/article/SB123413407376461353.html

[13] BWCI, "Indian language SMS." [Online]. Available: http://www.bwci.org.in/content.php?menu_id=2&sub_menu_id=6

[14] Wikipedia, "Standard Hindi." [Online]. Available: http://en.wikipedia.org/wiki/Standard_Hindi

[15] S. K. Kopparapu, I. A. Sheikh, and A. S. Pharande, "System and method for rapid prototyping of existing speech recognition solutions in different languages," Patent application number: 20100299133, 25 Nov 2010. [Online]. Available: http://www.faqs.org/patents/app/20100299133

[16] K. K. Bhuvanagiri and S. K. Kopparapu, "An approach to mixed language automatic speech recognition," in *Oriential COCOSDA*, 2010. [Online]. Available: http://desceco.org/O-COCOSDA2010/proceedings/paper_19.pdf

[17] Wikipedia, "Mixed language." [Online]. Available: http://en.wikipedia.org/wiki/Mixed_language

[18] C.-L. Huang and C.-H. Wu, "Generation of phonetic units for mixed-language speech recognition based on acoustic and contextual analysis," *IEEE Transactions on Computers*, vol. 56, pp. 1225–1233, 2007.

[19] P. Shih, J.-F. Wang, H.-P. Lee, H.-J. Kai, H.-T. Kao, and Y.-N. Lin, "Acoustic and phoneme modeling based on confusion matrix for ubiquitous mixed-language speech recognition," *Sensor Networks, Ubiquitous, and Trustworthy Computing, International Conference on*, vol. 0, pp. 500–506, 2008.

[20] AIR, "News broadcast." [Online]. Available: http://www.newsonair.com/writereaddata/broadcast/Hindi-Main-Bulletins-904.mp3

[21] ——, "News broadcast transcripts." [Online]. Available: http://www.newsonair.com/full_news.asp?type=bulletins&id=53

[22] CMU, "The cmu statistical language modeling (slm) toolkit." [Online]. Available: http://www.speech.cs.cmu.edu/SLM_info.html

[23] CRIS, "Indian railway." [Online]. Available: http://www.indianrail.gov.in