A Comparison of Microphone and Speech Recognition Engine Efficacy for Mobile Data Entry

Lumsden, Joanna; Durling, Scott; Kondratova, Irina

Publisher's version / Version de l'éditeur:

The International Workshop on Mobile and Networking Technologies for Social Applications (MONET'2008) [Proceedings], 2008

NRC Publications Archive Record / Notice des Archives des publications du CNRC :
https://nrc-publications.canada.ca/eng/view/object/?id=21dfe695-d543-4ded-bd63-c3be4c48541b
https://publications-cnrc.canada.ca/fra/voir/objet/?id=21dfe695-d543-4ded-bd63-c3be4c48541b

National Research Council Canada
Conseil national de recherches Canada

Institute for Information Technology
Institut de technologie de l'information

NRC · CNRC

*A Comparison of Microphone and Speech Recognition Engine Efficacy for Mobile Data Entry **

Lumsden, J., Durling, S., and Kondratova, I.
November 2008

* Proceedings of the International Workshop on MObile and NEtworking Technologies for social applications (MONET'2008), part of the LNCS OnTheMove (OTM) Federated Conferences and Workshops. November 9-14, 2008. Monterrey, Mexico, pp. 519-527.

Canada

A Comparison of Microphone and Speech Recognition Engine Efficacy for Mobile Data Entry

Joanna Lumsden, Scott Durling, and Irina Kondratova

National Research Council of Canada, IIT e-Business, 46 Dineen Drive, Fredericton, N.B., Canada E3B 9W4
{jo.lumsden, scott.durling, irina.kondratova}@nrc-cnrc.gc.ca

Abstract. The research presented in this paper is part of an ongoing investigation into how best to incorporate speech-based input within mobile data collection applications. In our previous work [1], we evaluated the ability of a single speech recognition engine to support accurate, *mobile*, speech-based data input. Here, we build on our previous research to compare the achievable speaker-*independent* accuracy rates of a variety of speech recognition engines; we also consider the relative effectiveness of different speech recognition engine and microphone pairings in terms of their ability to support accurate text entry under realistic mobile conditions of use. Our intent is to provide some initial empirical data derived from mobile, user-based evaluations to support technological decisions faced by developers of mobile applications that would benefit from, or require, speech-based data entry facilities.

Keywords: mobile speech input, microphone efficacy, speech recognition accuracy/efficacy, mobile technology, mobile evaluation.

1 Introduction

Although speech recognition has been nominated as a key potential interaction technique for use with mobile technologies [2-4], its widespread commercialization and adoption remains limited on account of unacceptable error rates [5]. It is estimated that accuracy rates can drop by as much as 20%-50% when speech is used in natural environments [3-5]. Achievable accuracy is a strong determinant of users' perception of speech recognition acceptability [2]; as such, it is important that we address the challenge of developing *effective* speech-based solutions for use in mobile settings.

A number of measures can be taken to increase recognition accuracy [2, 5-9]. The research presented in this paper focuses on two such measures: (1) empirically-based, context-specific selection of speech recognition engines (and microphones) to maximize potential accuracy within a given domain of use; and (2) identification of the effect of background noise and mobility on speech recognition accuracy. Specifically, we report on a comparison of the capacity of 5 different speech recognition engines to support accurate, mobile, speech-based text entry. The work discussed in this paper represents a continuation of our ongoing investigation in this

area [1]. Our previous work reported on an evaluation of the ability of a single speech recognition engine (SRE) to support accurate, mobile, speech-based data input; here, we build on our previous research to compare the achievable accuracy rates of a variety of SREs. In the following sections, we briefly describe the background to our work (Section 2) and the evaluation design and process (Section 3); we would refer interested readers to [1] for greater detail in both regards. In Section 4 we present, in detail, our results. In Section 5 we draw some *initial* conclusions from our observations.

2 Related Work

Speech recognition accuracy is typically degraded in noisy, mobile contexts because not only does background noise contaminate the speech signal received by the SRE but, additionally, people modify their speech under noisy conditions [5]. Specifically, in noisy environments, speakers exhibit a reflexive response known as the Lombard Effect [5, 10, 11] which causes them to modify the volume at which they speak and to hyperarticulate words. Since research suggests it is not possible to eliminate or selectively suppress Lombard speech, the onus is placed on SREs to be able to cope with variations in speech signals caused by mobile speech input under noisy and changing acoustic conditions [5, 10].

Under mobile conditions, background noise can confuse, contaminate, or even drown out a speech signal; as a result, SRE accuracy has been shown to steeply decline in even moderate noise [4, 5]. Even in stationary conditions, microphone type, placement, and quality affects user performance [12]; when the complexities of non-static usage environments are introduced, the influence of the microphone becomes even more pronounced [1, 5, 7-9]. Our previous research focused on assessing the impact of mobility and background noise on the efficacy of three different microphones [1] for supporting mobile speech-based data entry. Although previous research (see [1] for a detailed review) had been conducted into (a) the impact of mobility and background noise on speech recognition, and (b) the influence of microphone type on speech recognition, our prior work brought together, into one *novel* evaluation, many of the constituent – and previously unconnected – elements of previous studies to empirically compare the ability of three different microphones (appropriate in terms of form and function) to support accurate speech-based input under *realistic, mobile, noisy conditions*. In the research we present here, we extend the reach of our previous study to apply the input signals we recorded in our initial study to a series of SREs in order to compare their efficacy to support accurate speech-based input under realistic, mobile, noisy conditions.

3 Evaluation Design and Process

There are two components to describing our evaluation design and process: (a) the set-up from our previous study; and (b) the manner in which we extended the previous study to complete the research we report on here. In the following sections,

we provide a brief overview of the former (for extensive detail, see [1]) and then describe how we used the data collected in (a) to extend our analysis to compare multiple SREs.

3.1 Previous Experimental Design

Our previous study compared three commercially available and commonly used microphones – the NextLink Invisio Mobile (bone conduction) microphone [13], Shure's QuietSpot QSHI3 [14], and the Plantronics DSP-500 microphone [15] – in terms of their efficacy to facilitate mobile speech input. We developed a very simple data input application which ran on a tablet PC running Windows XP and used IBM's ViaVoice [16] speaker-independent SRE with a *push-to-talk* strategy. For each data entry item, participants were shown precisely what to enter, and given a maximum of three attempts in which to achieve an accurate entry. Participants were given training on how to use the system (in conditions commensurate with the actual experimental set-up) prior to commencing the study tasks.

We adopted a counterbalanced, between-groups design whereby participants were allocated to groups partitioned according to the three microphones; in counterbalanced order, each participant was required to complete a series of 10 data entry items under quiet environmental conditions, and 10 data entry items when surrounded by recorded city street sounds played at 70dB using a 7.1 surround sound system in our lab. While completing their data entry tasks, participants were required to be mobile using our 'hazard avoidance' or 'dynamic path system' – see [1] for more details.

Twenty four people participated in our study, 8 per microphone group. Since studies have shown that speech recognition rates are typically much lower for accented speakers [5], we restricted our recruitment to participants who were native English speakers with a Canadian accent; we recruited equal numbers of male and female participants, but restricted our age range to 18 – 35 year olds (young adults) to limit speaker variation that comes with age. In placing these restrictions on our participant recruitment, we recognize that we limited our ability to generalize from our results but we wanted to reduce the extraneous factors that *may* have impacted our results such that we were able to focus on the effects of the *microphones* rather than variances between the speakers; additionally, SREs available in North America are typically optimized to a 'generic' North American accent so by placing these restrictions on the participant recruitment we effectively tested the speech recognition technology within its self-proclaimed bounds.

Of the range of measures we recorded for analysis during our previous study, the data of interest to the study reported here is the series of speech signal (or voice) recordings upon which our SRE operated. In essence, each of these recordings captures what participants said, together with the background noise picked up by their respective microphones. In so doing, these recordings capture the presence of Lombard speech as well as the impact being mobile (i.e., walking) had on participants' speech: that is, the recordings are representative of likely speech patterns in real world contexts where users are exposed to noisy conditions as well as required

to be mobile while interacting with their mobile device. The following section describes how we utilized these recordings in our current study.

3.2 Current Study Design

As already mentioned, the intent of our current study was to extend our analysis to compare the efficacy of a range of SREs to that of the one used in our previous study (i.e., IBM's ViaVoice), as well as to determine if there was a single microphone-SRE pairing that proved to be most effective. Bearing in mind that we set up our previous experiment to reflect the real world context in which speech recognition may ultimately be used, and given that we took care to homogenize our participant group as far as possible with respect to accent and age, we feel that the results of our current investigation are valid as an *initial indication* of the potential benefits of one SRE over another for mobile interaction; that said, we suggest that the results of this study be considered as a baseline and acknowledge that the impact of speaker accent, especially, and age need to be investigated independently.

We compared the *speaker-independent* efficacy of (1) the IBM ViaVoice SRE (from our previous study) to the efficacy of 4 mainstream SREs: (2) the speaker-independent Sphinx 4 open source SRE (developed by CMU), set up with the default configuration and loaded with the specific grammars (as before) that were needed for the data entry application [17]; (3) Philips' SRE [18] and (4) Microsoft's SRE within Windows XP [19], both of which were set up with our required grammars; and (5) the Nuance Dragon Naturally Speaking SRE [20], which was essentially grammarless because it did not support grammar specification. Engines 3, 4, and 5 all support speaker-*dependence* (i.e., can be trained for specific users) but we used each in a speaker-*independent* mode in order to assess the 'walk-up-and-use' capabilities of each; we loaded a fresh profile for each participant such that the engines did not learn over time across participants.

We focused on first-attempt data input: that is, we filtered out, and only used, the recordings associated with participants' first attempts (successful or not) at inputting each data item. This decision not only allowed us to accommodate the fact that, for items correctly recognized first time by ViaVoice in our previous study, we only had one recording, but it also placed all our SREs on an equal footing.

We passed, using an automated process, the first-attempt voice recordings through each of our 4 additional SREs to derive a Boolean measure of accuracy for each data entry attempt. Our set of recordings included 160 files per microphone, 80 recorded under our quiet condition and 80 under our noisy condition. On this basis, each SRE was subjected to a total of 480 recordings, allowing us to test each against the three microphones and the two background audio conditions.

4 Results and Discussion

Our primary measure of accuracy was calculated as a ratio of the total number of first-attempt correct entries divided by the total number of tests (according to analytical breakdown). A multiple factor ANOVA showed that SRE ($F_{4,2370}$=27.03, $p<0.001$)

and the combination of SRE and microphone ($F_{8,2370}$=3.49, p=0.001) had a significant effect on accuracy. Figure 1 shows the accuracy rate achieved according to SRE. Tukey HSD tests showed that: the accuracy achieved using IBM ViaVoice was significantly higher than for all of the other engines; that the speech recognition accuracy achieved using Philips' SRE was significantly less than all of the other engines; and that the difference in accuracy achieved using the Sphinx, Microsoft, and Dragon engines was not statistically significant.

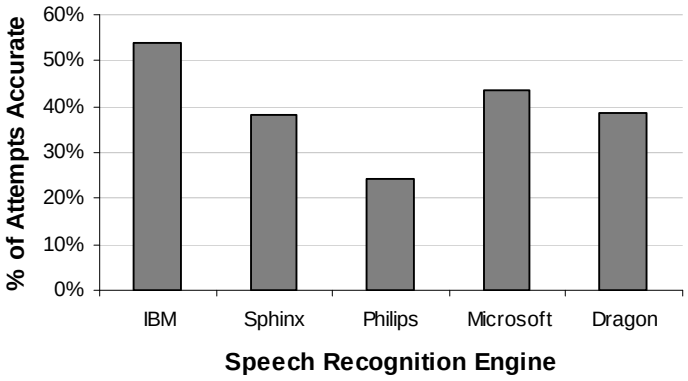

Fig. 1. Accuracy rate (accurate attempts/total attempts) according to speech recognition engine.

As can be seen from Figure 1, the maximum accuracy rate achieved on first attempt was approximately 54% for the IBM ViaVoice engine.

The results shown in Figure 1 are calculated irrespective of background noise. We did not find the combination of SRE and noise level to significantly affect accuracy ($F_{4,2370}$=0.42, p=0.796); furthermore, by analyzing the data across all background noise conditions, we obtain a picture of the likely average accuracy achievable by mobile or nomadic users who may typically move between noisy and quiet environments as they work.

Figure 2 shows the accuracy rates achieved according to SRE+microphone pairing. Tukey HSD tests showed that, with the exception of the Philips' SRE, when combined with the Invisio microphone, each SRE returned significantly lower accuracy rates than when combined with the other two microphones; there was no significant difference for these 4 SREs when combined with the QSHI3 and DSP-500 microphones. In the case of the Philips' SRE, however, the Invisio+SRE combination only returned significantly lower accuracy rates than the DSP-500+SRE combination (p<0.001); the Philips' SRE+QSHI3 and SRE+DSP-500 combinations did, however, return significantly different accuracy rates (p<0.001).

Focusing on the Invisio microphone across all 5 SREs, the Invisio+IBM and Invisio+Philips combinations returned significantly different accuracy rates (p<0.001), the former demonstrating a higher accuracy rate; the same was true for the Invisio+IBM and Invisio+Dragon combinations (p=0.01), for the Invisio+Microsoft and Invisio+Philips combinations (p<0.001), and for the Invisio+Microsoft and Invisio+Dragon combinations (p=0.01). When combined with the Invisio

microphone, there was no significant difference in the accuracy rates returned by the IBM, Sphinx, and Microsoft SREs.

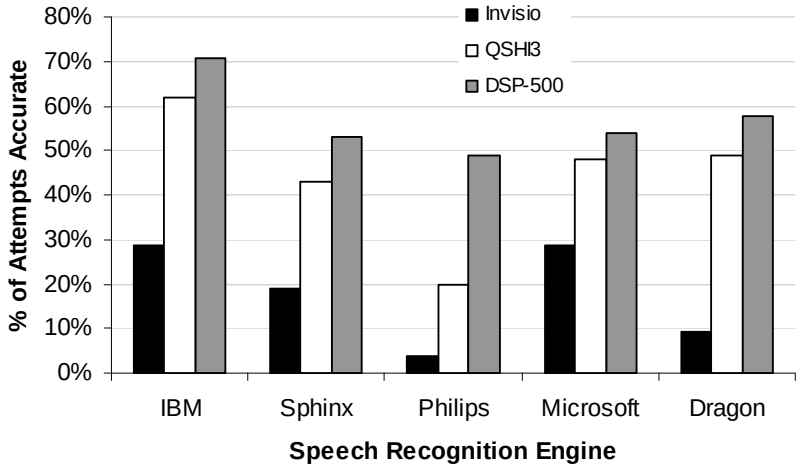

Fig. 2. Accuracy rate according to speech recognition engine and microphone.

The QSHI3+Philips combination was significantly less accurate than all of the other SREs combined with the same microphone ($p<0.001$ in each case); additionally, the QSHI3+IBM combination was significantly more accurate than the QSHI3+Sphinx combination ($p=0.016$). With these noted exceptions, there were no other significant differences for the QSHI3 microphone across the various SREs.

When combined with the IBM SRE, the DSP-500 microphone was significantly more accurate than when combined with the Sphinx ($p=0.04$) and Philips ($p<0.001$) SREs. There were no other significant comparisons across any of the other pairings for this particular microphone – most noticeably, unlike its performance with the other two microphones, the Philips SRE was on a par with the majority of the other SREs when combined with the DSP-500 microphone.

With the exception of the QSHI3+IBM, DSP-500+Microsoft, and DSP-500+Dragon combinations, the combination of DSP-500 microphone and IBM ViaVoice SRE returns the highest overall accuracy rates (approximately 71%). These results demonstrate the impact of pairing microphones with SREs to achieve the best possible potential for accurate speech recognition: for example, where ViaVoice's dominance is significantly reduced when paired with the Invisio microphone, the Philips' generally poor recognition is greatly boosted when paired with the DSP-500 microphone.

Of the total 952 accurately recognized data inputs, 41 were correctly recognized *despite* user error during input. We classify user error as instances where: users pressed the push-to-talk button but didn't speak (this was registered as an input attempt by the system); users pressed the push-to-talk button after they had started to speak or released it before they finished speaking (essentially clipping their recorded speech); or users simply said the wrong thing. Figure 3 shows the extent to which

each microphone+SRE coped with such errors to return a correct interpretation of the users' input.

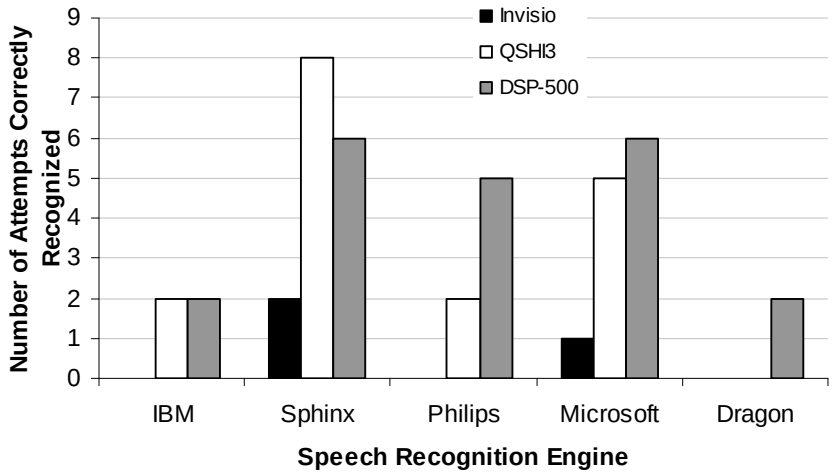

Fig. 3. Breakdown of correct recognition despite user input error.

Although we attribute no statistical significance to the tallies shown in Figure 3, it is interesting to note that certain microphone+SRE combinations appear better able to accommodate user input error. In particular, the fact that 12% of the correctly recognized flawed inputs are attributable to the Philips SRE+DSP-500 combination reflects, and perhaps accounts for, the significantly better accuracy rate achieved by this combination compared to the SRE's pairing with the other microphones (see Figure 2). Furthermore, although the IBM ViaVoice+DSP-500 combination has thus far excelled compared to the other SRE+microphone combinations, its dominance is much less when required to cope with flawed user inputs: in fact, the Sphinx+QSHI3 dominates in this capacity.

5 Conclusions & Preliminary Guidelines

Although it is premature to suggest concrete guidelines on the basis of our initial research, we close this paper with our conclusions and some preliminary guidance (which we call suggestions to reflect their current status) for designers.

Suggestion 1: Carefully consider the selection of microphone+SRE pairing, preferably by conducting empirical comparisons relative to intended context of use. Our results clearly indicate the importance of carefully considering SRE+microphone pairings relative to a specific context of use when developing speech-based mobile applications. By simply changing the microphone with which a given SRE is paired, it is possible to dramatically enhance the achievable accuracy rates – for example,

ViaVoice paired with the Invisio microphone returned a deplorable accuracy rate of 29% but this more than doubled to 71% when the same SRE was paired with the DSP-500 microphone. We have only compared microphone+SRE pairings relative to one context of use so are not in a position to make generalized recommendations concerning microphone+SRE pairings relative to various contexts of use at this time; furthermore, we would strongly encourage designers to conduct contextually-relevant, empirical analysis relative to the *specific* context for which they are designing a mobile application in order to elicit the most reliable data and thereby make the most informed decision.

Suggestion 2: Determine the likely extent to which target users will make mechanical or verbal errors during input (i.e., to what extent their physical environment and/or multitasking behavior may impact their capacity to devote attentional resources to speech-based input) and be prepared to trade off general accuracy against error tolerance. We have demonstrated the importance of considering, for any given application and domain, the extent to which users are likely to make mechanical or verbal errors during speech-based data entry. Our results suggest that designers may, depending on the context for which they are designing, have to make trade offs between SRE-microphone pairings that return high raw accuracy and pairings that have an increased ability to cope with flawed input.

Suggestion 3: Carefully consider the requirement for accurate first time data entry versus scope to tolerate repeated entries in order to enter data correctly. Our study only looked at first attempt accuracy; while this is often essential, we recognize that under situations where multiple attempts to achieve an accurate input would be tolerable, the breakdown of ultimate accuracy rates across the SREs we tested might differ.

Suggestion 4: Carefully consider the applicability of different microphone designs relative to the intended context of use. We recommend that designers carefully consider not only the accuracy that can be achieved using a given microphone, but also its appropriateness to the context in which it is to be used – e.g., if a user has to wear a safety helmet or to use specific equipment such as a stethoscope, to what extent can a given headphone mounted microphone be accommodated or does and alternative form factor need to be sought? Accuracy alone is insufficient to make the microphone usable.

Finally, as previously discussed, our study is not without its limitations; as such, we present our results within the scope of our noted caveats. We were testing speaker-*independent* operation of the SREs (since our interest was in their 'walk-up-and-usability'); we recognize that if the speaker-*dependent* SREs (Philips, Microsoft, and Dragon) were to be trained, their accuracy rates would likely dramatically improve. That being said, our results not only demonstrate the difference in the capabilities of these systems (which are normally trained prior to use) to cope with speaker-*independent*, walk-up-and-use situations, but we also present the results as empirical data to assist a designer when selecting an SRE and microphone for use in a speaker-*independent* capacity. At the very least, we have empirically highlighted the complexity of decisions surrounding microphones and SREs for mobile applications; we have provided data that was not previously available to designers and, as such, hope that it not only proves useful to designers of speech-based mobile data input, but

also highlights those areas that require detailed consideration when making speech technology decisions during the design process.

References

1. Lumsden, J., Kondratova, I., and Durling, S., Investigating Microphone Efficacy for Facilitation of Mobile Speech-Based Data Entry, In: Proceedings of British HCI'2007 Conference, pp. 89-98, Lancaster, UK, 3-7 Sept, (2007).
2. Price, K., Lin, M., Feng, J., Goldman, R., Sears, A., and Jacko, J., Data Entry on the Move: An Examination of Nomadic Speech-Based Text Entry, In: Proceedings of 8th ERCIM Workshop "User Interfaces For All" (UI4All'04), pp. 460-471, Vienna, Austria, 28-29 June, (2004).
3. Sawhney, N. and Schmandt, C., Nomadic Radio: Speech and Audio Interaction for Contextual Messaging in Nomadic Environments, *ACM Transactions on Computer-Human Interaction*, ***7***(3), pp. 353 - 383, (2000).
4. Ward, K. and Novick, D., Hands-Free Documentation, In: Proceedings of 21st Annual International Conference on Documentation (SIGDoc'03), pp. 147 - 154, San Francisco, USA, 12 - 15 October, (2003).
5. Oviatt, S., Taming Recognition Errors with a Multimodal Interface, *Communications of the ACM*, ***43***(9), pp. 45 - 51, (2000).
6. Lumsden, J., Kondratova, I., and Langton, N., Bringing A Construction Site Into The Lab: A Context-Relevant Lab-Based Evaluation Of A Multimodal Mobile Application, In: Proceedings of 1st International Workshop on Multimodal and Pervasive Services (MAPS'2006), pp. 62 - 68, Lyon, France, June 29, (2006).
7. Sammon, M., Brotman, L., Peebles, E., and Seligmann, D., MACCS: Enabling Communications for Mobile Workers within Healthcare Environments, In: Proceedings of 8th International Conference on Human Computer Interaction with Mobile Devices and Services (MobileHCI'06), pp. 41 - 44, Helsinki, Finland, 12 - 15 September, (2006).
8. Sebastian, D., Development of a Field-Deployable Voice-Controlled Ultrasound Scanner System, M.Sc. Thesis, Dept. of Electrical and Computer Engineering, Worcester Polytechnic Institute, Worcester, MA, USA, (2004).
9. Vinciguerra, B., A Comparison of Commercial Speech Recognition Components for Use with the Project54 System, M.Sc. Thesis, Dept. of Electrical Engineering, University of New Hampshire, Durham, NH, USA, (2002).
10. Pick, H., Siegel, G., Fox, P., Garber, S., and Kearney, J., Inhibiting the Lombard Effect, *Journal of the Acoustical Society of America*, ***85***(2), pp. 894 - 900, (1989).
11. Rollins, A., Speech Recognition and Manner of Speaking in Noise and in Quiet, In: Proceedings of Conference on Human Factors in Computing Systems (CHI'85), pp. 197 - 199, San Francisco, USA, 14 - 18 April, (1985).
12. Chang, J., Speech Recognition System Robustness to Microphone Variations, M.Sc. Thesis, Dept. of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, MA, USA, (1995).
13. NextLink, Invisio Pro, http://www.nextlink.se/.
14. Shure, QuietSpot QSHI3, http://www.sfm.ca/quietspot/qshi3.html.
15. Plantronics, DSP-500 Headset, http://www.plantronics.com/north_america/en_US/products/cat640035/cat1430032/prod440044.
16. IBM, Embedded ViaVoice, http://www-306.ibm.com/software/pervasive/embedded_viavoice/.
17. CMU, Sphinx-4, http://cmusphinx.sourceforge.net/sphinx4/.

18. Philips, Speech SDK, http://www.speechrecognition.philips.com/index.asp?id=521.
19. Microsoft, Windows Desktop Speech Technology, http://msdn.microsoft.com/en-us/library/system.speech.recognition.aspx.
20. Nuance, Dragon Naturally Speaking, http://www.nuance.com/naturallyspeaking/sdk/client/.