IDIAP RESEARCH REPORT

RESEARCH INSTITUTE

FROM RESEARCH TO REALITY: EVALUATION OF A SINGLE-COMPUTER REAL-TIME LVCSR SYSTEM FOR SPEECH-BASED RETRIEVAL

Andrei Popescu-Belis Maryam Habibi
Philip N. Garner Nan Li

Idiap-RR-12-2017

MARCH 2017

Centre du Parc, Rue Marconi 19, P.O. Box 592, CH - 1920 Martigny
T +41 27 721 77 11 F +41 27 721 77 12 info@idiap.ch www.idiap.ch

FROM RESEARCH TO REALITY: EVALUATION OF A SINGLE-COMPUTER REAL-TIME LVCSR SYSTEM FOR SPEECH-BASED RETRIEVAL

A. Popescu-Belis, M. Habibi, P. N. Garner

Idiap Research Institute
Centre du Parc, Rue Marconi 19
CH-1920 Martigny, Switzerland

Nan Li[*]

Ecole Polytechnique Fédérale, Lausanne (EPFL)
Pedagogical Research and Support (CRAFT)
CH-1015 Lausanne, Switzerland

ABSTRACT

This paper presents a series of tests that were performed on a state-of-the-art real-time automatic speech recognition system for English, in a single-computer implementation. As the intention is to use the system for speech-based query-free document retrieval in conversations, several parameters were varied: text type, microphone quality, computing power, speaker fluency, and pace of the speech. Word accuracy over various word counts, including a restriction to content words, varied in the 30%–70% range. The paper compares results over many conditions, and concludes that the ASR system is acceptable for the intended use only if all the parameters are in optimal conditions. If more than two parameters are suboptimal, then its output becomes too noisy for document retrieval.

Index Terms— Automatic speech recognition, audio user interfaces, human factors, performance testing

1. INTRODUCTION AND MOTIVATION

The performance of automatic speech recognition (ASR) systems can be evaluated intrinsically, often in terms of word error rate in specific test conditions, or extrinsically, in terms of the utility of ASR to the application in which it is embedded. In the latter case, evaluation can lead to a go/no-go decision ("is the system usable at all for our application?") or to gradual judgments, which are most meaningful when several systems are compared ("should we use System A or System B for our application?").

In this paper, we present our method and results for deciding whether, or to what extent, a real-time, single-computer, large-vocabulary continuous ASR system can be used for speech-based document recommendation in conversations. More specifically, the intended application is the Automatic Content Linking Device (ACLD, see [1, 2]), a just-in-time retrieval system which uses the words recognized through ASR within a conversation in English, typically a business meeting, to perform spontaneous searches and to recommend potentially relevant documents to meeting participants. In such a setting, the role of the ASR system is essential, although a certain percentage of errors can be tolerated, as techniques exist to mitigate their impact on the subsequent searches performed by the ACLD.

The evaluation experiment presented here aims to assess whether a research ASR system available to us [3, 4] fulfills the needs for use in the ACLD, depending on a number of contextual parameters and system settings, some of which can be controlled and some others which cannot. The challenge of the evaluation is to compare over a number of parameters without testing individually each and every possible combination – which is impractical both for the human subjects and for the evaluators who have to deal with a very large number of scores, blurring the global picture. Our experiments focused on the variation of the following parameters: (1) computing power, (2) microphone and sound card, (3) speakers' fluency, (4) text type (monologue or dialogue), (5) pace of speech (normal or paused), and (6) presence or absence of noise (from a tabletop projector). The goal was to reach a go/no-go decision in a given setting.

This paper is organized as follows. In Section 2 we briefly present the ASR system. In Section 3 we present the setup of the experiments, giving more details about the parameters that were tested. The observations gathered from the actual execution of the evaluations are summarized in Section 4, and the evaluation metrics that are used are listed in Section 5. The results shown in Section 6 are discussed in Section 6, leading to the conclusion from the main experiment, and a comparison with other evaluations of the ACLD.

2. THE ASR SYSTEM

For the ACLD [1, 2], we use a real-time, speaker-independent, large-vocabulary ASR system developed by the AMI Consortium [3, 4]. The system used here has a dictionary of 50,000 words, including general vocabulary and vocabulary specific

[*] The four authors are grateful to the Swiss National Science Foundation (SNSF) for its support through the IM2 NCCR on Interactive Multimodal Information Management (see http://www.im2.ch). The authors would like to thank Omar Mubin and Roland Tormay for their participation in the experiments, and Pierre Dillenbourg for overall support.

to the AMI Meeting Corpus [5]. One of the main features is the use of a pre-compiled grammar, which allows it to retain accuracy even when running in real-time on a low resource machine. The ASR system can also run slower than real-time, to maximize accuracy of recognition, but the gain over the real-time mode is small (about 1%).

For the RT07 meeting data (see [6]), when using signals from individual headset microphones, the AMI ASR system reaches about 38% word error rate. With a microphone array, this increases to about 41%. These values indicate that, in theory, the ASR could sense enough correct words to make it applicable to the ACLD – a claim that we attempt to validate through the experiments presented in this paper.

To make optimal use of the ASR, additional software is necessary to feed it data from a continuous audio stream. A voice activity detector (VAD) segments the signal into speech and non-speech (noise or silence), based on its mean energy, and conveys only the speech segments to the ASR. The VAD continuously attempts to estimate the mean level of noise, and labels as speech an audio signal that exceeds noise (by a certain threshold) for a period longer than a certain time. If the signal is below the noise threshold, again for a period longer than a certain time, the VAD considers it as non-speech.

The noise threshold, i.e. the energy level above noise above which a signal is considered as speech, was set at 10 dB (the dynamic range of conversational speech is 20–30 dB). This parameter can be increased for noisy environments, especially with irregular noise, or decreased slightly for silent environments, or for very directional or low-noise microphones. The length of the time intervals in which the signal needs to be above (resp. below) threshold so that the VAD switches its state from silence to speech (resp. vice-versa) were set to 0.300 s, close to the lowest limit, to force the system to be more sensitive to pauses and thus segment utterances more often.

3. SETUP OF THE EXPERIMENTS

The intention of the ASR evaluation experiments is to estimate the accuracy of the ASR system in a variety of conditions, which could all potentially occur during the intended use of the ACLD system. Many parameters, already mentioned in the introduction, could thus vary, and while we do not expect high ASR performance in all of the conditions, a low performance in a majority of situations would speak against the use of the system.

We first provide here a bird-eye view of the parameters (see Table 1 for a meaningful representation). Experiments were carried out in two locations, at the Idiap Research Institute and the CRAFT laboratory at EPFL. One characteristic of the location was the **computer** that was used to run the ASR and some elements of the ACLD, which was a high-end laptop at Idiap, and an entry-level desktop computer at CRAFT (see Section 3.1). The end-user intended setting at CRAFT included a tabletop projection device called TinkerLamp [7] to allow users to interact with the recommended documents; however, this was also a big source of **noise**, so we tested conditions with the lamp turned off as well. Both rooms were otherwise silent. The **audio hardware** could be moved from one location to the other, and included a high-end headset and sound card, a commercial microphone array, and a low-end USB headset (see Section 3.2). The subjects were asked to read a text, or to enact a dialogue in pairs, thus varying the **text type**. Their **fluency** varied from low to high, and in half of the conditions they were asked to lower their **pace** by pausing between each sentence. Otherwise, they were instructed to read aloud at their natural speech pace.

3.1. Hardware: Computers

The constraints of the ACLD required the use of a non-distributed ASR system, mainly for portability to various meeting rooms. In any case, the ASR system was not parallelized, therefore decoding on one processor core at a time, even when multi-core processors are available. However, automatic load balancing from the OS ensures that, in general, other demanding processes are run on other cores if available. The ASR system was originally developed for Linux, and then ported to Mac OS.

Two Macintosh computers, one from Idiap and one from CRAFT, were used in the tests: respectively a MacBook Pro laptop and a Mac Mini small desktop computer. Both had Mac OS X 10.6 (Snow Leopard) with a 64-bit kernel. The Mac Mini had a Server version of the OS, with no impact on our tests. The MacBook Pro (model 8,1 from May 2011) had a dual-core 2.7 GHz Intel Core i7 processor, with 4 MB on-chip L3 cache, and 8 GB of SDRAM. The Mac Mini (model 4,1 from July 2010) had an Intel Core 2 Duo P8800 processor at 2.66 GHz, with 3 MB on-chip L2 cache, and 4 GB of SDRAM. The main differences are thus the improved processor and especially the larger memory size (8 GB vs. 4 GB) of the MacBook Pro.

3.2. Hardware: Microphones

The ACLD monitors the conversation in a meeting, therefore, in the ideal case, a far-field tabletop microphone should be used. However, in preliminary tests, the performance of such microphones appeared to be unsatisfactory, hence in these experiments we use either head-worn individual microphones, or a microphone array. The two rather high-end microphones were coupled to an external sound card, but a low-end USB head-worn microphone was also used. Here are their specifications:

1. Microphone array: Microcone by dev/audio, a 7-channel array with software for Mac OS X, designed to capture group conversations, locate the current speaker,

Text type	Computer	Microphone	Fluency	Pace	Noise	reading time (s)
Dialogue	Mac Mini	High-end headset	2+2	Normal	On	150
				Paused	On	240
			3+1	Normal	Off	165
				Paused	Off	210
		Microphone array	3+1	Normal	Off	165
				Paused	Off	210
	MacBook Pro	Microphone array	3+1	Normal	–	195
				Paused	–	215
Text	Mac Mini	High-end headset	2	Normal	On	*240
				Paused	On	300
	MacBook Pro	Low-end headset	3	Normal	–	80
				Paused	–	120
			1	Normal	–	135
				Paused	–	150
		Microphone array	3	Normal		95
				Paused		120

Table 1. Conditions of the evaluation experiments, with parameters organized in a meaningful order. Fluency in English is coded on a 3-point scale (1 being lowest). Noise comes from a projector lamp on the meeting table at CRAFT. The duration (in seconds) is for reading the texts, '*' denotes an experiment in which the ASR did not complete recognition.

and perform beamforming on the audio signal to enhance it [8]. The Microcone has a sample rate of 48 KHz, with 24-bit resolution, and a frequency response of 20–20,000 Hz.

2. High-end microphones (Sept. 2011): (1) Shure SM10A, a head-worn unidirectional dynamic microphone with a cardioid pattern and a 50–15,000 Hz frequency response, often used for ASR. (2) AKG C520, an electret but otherwise similar microphone with a 60–20,000 Hz frequency range. The microphones were plugged using their XLR connectors into the following sound card, with phantom power for the AKG one.

3. Sound card: Presonus FireStudio, a standalone 24-bit 96 kHz recording system with eight microphone pre-amplifiers and a FireWire output of the digitally-converted signal. The signals of the two microphones were mixed before being conveyed to the ASR system.

4. Low-end microphone: Logitech H555 headset (May 2011) with USB microphone, frequency response of 100–10,000 Hz, plugged into the USB port. Several such devices could be mixed using a USB port replicator, but given the foreseen performance decrease, the low-end microphone was only used for monologues.

The input volume for each microphone was set by recording some input from the microphone and listening back to it, trying to avoid saturation. A low perceived volume is acceptable if the voice is clearly distinguished from noise. For the external sound card, the best setting is at the maximum input and output levels.

3.3. Texts and Speakers

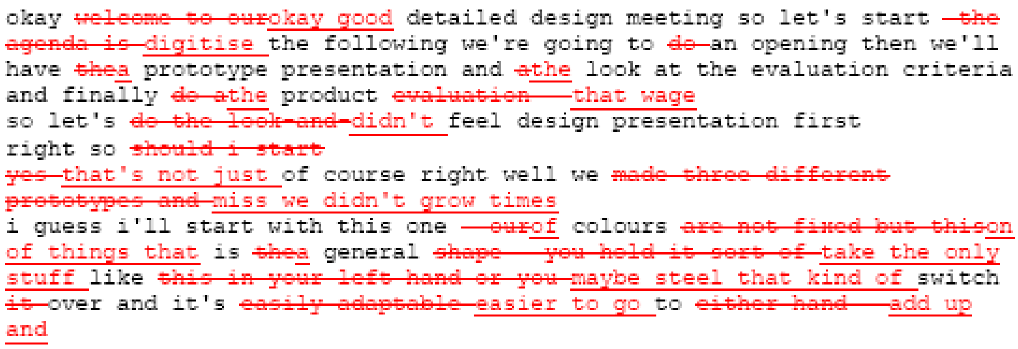
We selected two texts, a dialogue and a monologue, to be read aloud by participants to the experiment. Of course, this is quite unlike many ASR evaluation methods, which make use of spontaneous speech, and use a post-hoc transcript of what was said as a reference against which ASR output is compared. By using written texts we spare the effort of providing an exact post-hoc transcription, although we might ease the task of the ASR through cleaner input.

The monologue is an article from the New York Times Sunday Magazine [9], with an interview of a physicist, in written form, hence closer to the spoken form than a regular article. We used only a fragment of 247 words from its beginning. The dialogue is an excerpt of transcript ES2008d from the AMI Meeting Corpus [5], about the design of a remote control, which was edited to feature only two speakers and remove some disfluencies. A fragment of 486 words was selected. The AMI ASR system was trained on the AMI corpus, which helps to have fewer out-of-vocabulary words.

Five speakers took part in our experiments, and their estimated fluency (related of course to their proficiency, but also including "ASR-friendliness") is coded on a 3-point scale, from native or very fluent (3), to understandable but not-so-fluent (1). In some of the dialogue experiments, we mixed two participants with opposite fluencies (1+3).

4. EXECUTION OF THE EVALUATION

At the start of each experiment, the ASR system was first turned on, using the ACLD control panel. Once the real-time recognition was running (as tested by uttering "one, two,



okay ~~welcome to our~~okay good detailed design meeting so let's start ~~the~~
~~agenda is~~ digitise the following we're going to ~~do~~ an opening then we'll
have ~~the~~a prototype presentation and ~~a~~the look at the evaluation criteria
and finally ~~do a~~the product ~~evaluation~~ that wage
so let's ~~do the look and~~didn't feel design presentation first
right so ~~should i start~~
~~yes~~ that's not just of course right well we ~~made three different~~
~~prototypes and~~ miss we didn't grow times
i guess i'll start with this one ~~our~~of colours ~~are not fixed but this~~on
of things that is ~~the~~a general ~~shape you hold it sort of~~ take the only
stuff like ~~this in your left hand or you~~ maybe steel that kind of switch
~~it~~ over and it's ~~easily adaptable~~ easier to go to ~~either hand~~ add up
and

Fig. 1. Differences between ASR output (underlined) and reference text (strikethrough) for the beginning of the AMI dialogue. Condition: MacBook Pro, microphone array, mixed fluency (high + low), and normal speech pace. The effects of fluency are visible when comparing the first lines with the last ones.

three" for instance), the subject(s) started reading their text(s) – see reading times in Table 1. The ASR output from a log window was saved into a text file, and any particular behavior of the ASR was noted. Comparisons with the ground truth were performed later, based on the text files. Speakers were not recorded, as we do not plan to analyze their production: obviously, for the low-fluency speakers, their speech contained a number of disfluencies that degraded ASR performance.

In several cases, the ASR system appeared to lag increasingly behind the speakers. This happened on the Mac Mini when a fluent speaker read a text in normal speed, while slower speaker did not experience this problem. In one case reported here (line 9 of Table 1), the delay increased to the point where the ASR system stopped producing output. (In fact, in this case and a neighboring one, see lines 9 and 10 in Table 2, speakers read more than the 247-word fragment of the other experiments.) Although this is also an important parameter, we did not investigate it any further in these experiments, because evidence from numerous long demonstrations (up to one hour) in non-adverse conditions showed that this was an infrequent problem.

As seen from Table 1, not all possible combinations of parameters were tested, and not all are reported here, otherwise a very large number of sessions would have been necessary. The guiding question was: what are the main parameters that have a determining influence on the ASR output, and how robust are they with respect to unfavorable values?

5. SCORING THE ASR OUTPUT

To evaluate the accuracy of the ASR output (1-best decoding result) in comparison with the read texts – excluding production mistakes – we consider word accuracy scores over several subsets of words: (1) all (raw) words, (2) all words converted to singular form, (3) stemmed words, (4) all words minus stop words, and (5) selected content words. While the first options correspond to traditional word accuracy, the last one reflects more closely the needs of our application, which uses mainly content words for document retrieval. In the Section 6 we will focus on the first and last counts. The reference text was lower-cased to match the ASR output, its punctuation was removed, and some differences between British and US spelling were solved. These matters might also be considered for evaluation in further experiments.

For the quantitative measures of word accuracy, we use the following traditional notations: N is the number of words in the *reference texts*, H is the number of *correct* labels hypothesized by the ASR system in a given experiment, D the number of deletions, S the number of substitutions, and I the number of insertions. The definition of *correctness* is $C = H/N$ and word accuracy is $\mathit{WAcc} = (H - I)/N$. Note that word error rate is $\mathit{WER} = (S + D + I)/N$ and $\mathit{WAcc} + \mathit{WER} = 1$ because $H + S + D = N$. We use the implementations provided by HTK version3.4.1.

When only content words are counted, these are first marked (once and for all) in the reference file, stemmed, and put into a reference list (66 and 176 words respectively). This list is then used to extract occurrences of content words from the candidate ASR output, which are also stemmed and put into a list. Then both lists are submitted to the WER scoring software. Though this procedure could be improved, we consider it an efficient way to count ASR accuracy for a limited set of content words, without other human intervention than annotating them in the reference text.

A simple but useful visualization of the accuracy of the ASR output can be obtained using the Track Changes functionality of the Microsoft Word text processing software, as shown in Figure 1. This makes visible the insertions, deletions, and spans of correctly recognized text, at the word level, using an edit distance algorithm.

Type	Computer	Microphone	Prof.	P/L	C	WAcc	H	D	S	I	N
Dialogue	Mac Mini	High-end h.	2+2	N⋆	.38	.33	185	105	196	26	486
				P⋆	.58	.51	280	41	165	32	486
			3+1	N	.28	.27	136	107	243	4	486
				P	.42	.33	204	68	214	43	486
		Mic. array	3+1	N	.54	.49	262	45	179	23	486
				P	.53	.48	260	35	191	28	486
	MacBook Pro	Mic. array	3+1	N	.56	.50	270	45	171	29	486
				P	.56	.48	273	47	166	40	486
Text	Mac Mini	High-end h.	2	N⋆	.30	.26	177	104	303	26	*584
				P⋆	.53	.45	380	63	279	56	*722
	MacBook Pro	Low-end h.	3	N	.72	.66	177	11	59	14	247
				P	.74	.66	184	5	58	22	247
			1	N	.35	-.08	86	3	158	106	247
				P	.40	.07	98	8	141	81	247
		Mic. array	3	N	.62	.57	153	15	79	13	247
				P	.57	.51	141	15	91	15	247

Table 2. Correctness, word accuracy, and raw word counts (correctly hypothesized, deletions, substitutions, insertions, and reference) for a subset of conditions presented in a meaningful order. 'P/L' indicates speech pace (normal or paused) and presence of noise ('⋆' for lamp on). In lines 9 and 10 (marked with a '*'), the speaker read beyond the 247-word limit.

Condition					Word accuracy						
Type	**Computer**	**Microphone**	**Prof.**	**P/L**	**Raw**	*P-N*	**Sing.**	**Stem**	**Stop**	**Cont.**	*P-N*
Dialogue	Mac Mini	High-end h.	2+2	N⋆	33	*+18*	33	33	26	35	*+22*
			3+1	N	27	*+6*	28	28	22	34	*+10*
		Mic. array	3+1	N	49	*-1*	49	49	48	53	*-1*
	MacBook Pro	Mic. array	3+1	N	50	*-2*	51	52	45	61	*-3*
Text	Mac Mini	High-end h.	2	N⋆	26	*+19*	26	27	26	36	*+17*
	MacBook Pro	Low-end h.	3	N	66	*0*	68	69	67	74	*0*
			1	N	-8	*+15*	-5	-4	7	18	*+9*
		Mic. array	3	N	57	*-6*	59	59	56	53	*+5*

Table 3. Word accuracy scores (percentages) for five word count methods: all words (raw), singular form, stemmed words, stop words removed, and content words only. The changes in scores for paused speech with respect to normal pace are noted '*P-N*'.

6. RESULTS AND DISCUSSION

For a start, a visualization example is shown in Figure 1 for the AMI dialogue. While the display does not provide a quantitative measure of accuracy, it provides an overall impression of the usability of the output. The initial part, read by a highly fluent speaker, is very accurate, with almost no content word missed (apart from 'evaluation'). However, the second part (from "right so should I start"), read by a low-fluency speaker, has a majority of words wrong, although some useful content words are recognized ('colours', 'switch').

Table 2 shows the values of the word counts, along with Correctness and Word Accuracy scores, when all words are used for scoring. Although we have similar tables for the other ways of counting, we focus the discussion on overall results, shown as percentages in Table 3 for the five ways of counting words. In fact, the values and variations of scores across conditions are similar when counting all raw words vs. singular forms vs. stems. When removing stopwords from raw words, the variation is also similar, although the scores are lower, as ASR performance on stopwords seems better than average.

The main finding is that there is a lot of variability: 27%–66% word accuracy, and 30%–74% for content words. The best condition is for a high-fluency speaker reading the monologue, on the more powerful computer, regardless of the microphone but on condition that there is no external noise. In fact, the best accuracy (66%, and 74% for concepts only) is reached with the low-end headset. The Microcone appears to provide good results too, even with the Mac Mini, but only if there is no noise and with a fluent speaker (57% accuracy). Non-fluent speakers score generally low: e.g. the close-to-zero accuracy on the monologue, with the MacBook Pro and the low-end headset. The text type did not have a strong impact on scores.

The most important factors are thus the *fluency* of the speakers, over which one has limited control once the system is deployed, and the *computation power*, which, although restricted to one computer as the ASR system is not distributed, can be increased by the purchase of high-end computers, with a high-end dual-core processor and at least 8 GB of RAM. The use of load-balancing software for a multi-core processor could ensure that ASR decoding gets one core at 100%.

The *audio devices* do not seem to have an essential impact. Our tests showed that the Microcone is perfectly suitable for conversational ASR, and in particular it gave acceptable scores even on the low-performance computer, with fluent speakers and/or paused speech. However, the Microcone did not work with a significant noise source in the background (the projector) even when manually forcing the beamforming software to exclude the direction of the noise source. The low-end headset had very satisfactory scores for fluent speakers, and was about 10 times cheaper than the high-end setting (two microphones and a sound card), but using more than one USB headset is not convenient. Note that the cost of the high-end setting, with two microphones, was only about half of the Microcone.

There is no need to pause between sentences if computing power is available. The conditions with pauses between sentences are not much better than those with normal speech, on condition that the computer can keep up the pace, which is not the case for the Mac Mini, for which paused speech (though unnatural) greatly helps the ASR.

Overall, fluent speakers, low noise, a powerful computer and high-end microphones or the Microcone lead to highly acceptable recognition results. However, if two or more of these requirements are not met, then accuracy decreases quite dramatically from around 70% to around 40%, or even less if more requirements are missing. A correct setting of the parameters for the ASR and the VAD is essential too. Therefore, in the setting intended at CRAFT, where these conditions were not met, *the ASR system was not suitable for the ACLD*. Alternative solutions, e.g. based on the exploitation of hand-written notes, are currently explored.

However, using a high-end computer with the Microcone in an environment of reasonably fluent English speakers brings the ASR output to usable levels for the ACLD. Of course, given the state-of-the-art in ASR, it is not likely that an alternative ASR system would increase performance in this setting, with multi-party, continuous, conversational speech. Therefore, we will work on the improvement of system parameters (VAD, dictionaries) as well as post-ASR processing of the word.

This paper has presented evaluation results for a state-of-the-art ASR system intended to be used within a document recommendation system for conversations. While most of the low-performing conditions could be improved through more technical work, our results illustrate the difficulties of deploying a system in a real-life context from an ASR user-oriented perspective. The importance of higher computing power and more fluent speakers was demonstrated, corroborating common knowledge in the field. More innovatively, we brought evidence that microphone quality was not essential, and that a novel microphone array was very suitable, while the pace of speech can be kept natural for a long time if computing power is adequate.

7. REFERENCES

[1] Andrei Popescu-Belis and al., "The AMIDA Automatic Content Linking Device: Just-in-time document retrieval in meetings," in *Proceedings of MLMI 2008*, LNCS 5237, pp. 272–283. Utrecht, 2008.

[2] Andrei Popescu-Belis, Majid Yazdani, Alexandre Nanchen, and Philip N. Garner, "A speech-based just-in-time retrieval system using semantic search," in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, 2011, pp. 80–85.

[3] Philip N. Garner, John Dines, Thomas Hain, Asmaa El Hannani, Martin Karafiat, Danil Korchagin, Mike Lincoln, Vincent Wan, and Le Zhang, "Real-time ASR from meetings," in *Interspeech 2009 (10th Annual Conference of the Intl. Speech Communication Association)*, Brighton, UK, 2009, pp. 2119–2122.

[4] Thomas Hain and Philip N. Garner, "Speech recognition," in *Multimodal Signal Processing: Human Interactions in Meetings*, Steve Renals, Hervé Bourlard, Jean Carletta, and Andrei Popescu-Belis, Eds., pp. 56–83. Cambridge University Press, Cambridge, UK, 2012.

[5] Jean Carletta, "Unleashing the killer corpus: experiences in creating the multi-everything AMI Meeting Corpus," *Language Resources and Evaluation Journal*, vol. 41, no. 2, pp. 181–190, 2007.

[6] Rainer Stiefelhagen, Rachel Bowers, and Jonathan Fiscus, Eds., *Proceedings of CLEAR 2007 and RT 2007*, LNCS 4625. Baltimore, MD, 2008.

[7] Guillaume Zufferey, Patrick Jermann, and Pierre Dillenbourg, "A tabletop learning environment for logistics assistants: activating teachers," in *Proceedings of the 3rd IASTED Int. Conf. on Human Computer Interaction (HCI'08)*, Innsbruck, 2012, pp. 37–42.

[8] Iain McCowan, "Microphone arrays and beamforming," in *Multimodal Signal Processing: Human Interactions in Meetings*, Steve Renals, Hervé Bourlard, Jean Carletta, and Andrei Popescu-Belis, Eds., pp. 28–39. Cambridge University Press, Cambridge, UK, 2012.

[9] Deborah Solomon, "Greene, with curiosity: Questions for Brian Greene," *The New York Times Sunday Magazine*, pp. 16, December 19, 2010.