

The design and collection of COSINE, a multi-microphone in-situ speech corpus recorded in noisy environments

Alex Stupakov^a, Evan Hanusa^a, Deepak Vijaywargi^a, Jeff Bilmes^{*,a}, Dieter Fox^b

^a*Department of Electrical Engineering, University of Washington, Seattle, WA, USA*

^b*Department of Computer Science & Engineering, University of Washington, Seattle, WA, USA*

Abstract

We present an overview of the data collection and transcription efforts for the COⁿversational Speech In Noisy Environments (COSINE) corpus. The corpus is a set of multi-party conversations recorded in real world environments, with background noise, that can be used to train noise-robust speech recognition systems or develop speech de-noising algorithms. We explain the motivation for creating such a corpus, and describe the resulting audio recordings and transcriptions that comprise the corpus. These high quality recordings were captured in-situ on a custom wearable recording system, whose design and construction is also described. On separate synchronized audio channels, seven-channel audio is captured with a 4-channel far-field microphone array, along with a close-talking, a monophonic far-field, and a throat microphone. This corpus thus creates many possibilities for speech algorithm research.

Key words: Speech recognition, microphone array, multi-party, multi-microphone, portable recording, noise-robust speech recognition

1. Introduction

In many applications, practical automatic speech recognition (ASR) systems must be robust to the presence of background noise in the environment. These applications are numerous, and include dictation software; speech-based human-computer interfaces; speech recognition of telephone or air traffic control conversations; speech recognition or keyword search of TV or radio programs; and voice commands used by automobile drivers, soldiers, firefighters, law enforcement officials, or disabled individuals to interact with assistive devices.

Two types of effects must be overcome when training a speech recognition system that must work in noisy environments: the presence of additive or convolutional noise as well as reverberation, and the effects of the noisy environment on the nature of the speech (e.g., the Lombard effect). The methods used to mitigate these effects fall into three main categories [1]:

1. Perform noise cancellation or reduction on the audio signal before passing it into the speech recognizer.
2. Use noise-robust feature extraction methods to gain performance improvements over standard MFCC features [2]. For example, mean/variance normalization, feature smoothing [3], and a variety of other feature cleaning/enhancement techniques show improvement over standard MFCC/PLP features [4, 5, 6].
3. Train the acoustic model on a combination of clean and noisy speech, or use speech recorded in the desired noisy environment to adapt an existing acoustic model. The performance of a system trained

*Corresponding author

Email addresses: stupakov@ee.washington.edu (Alex Stupakov), hanusaem@ee.washington.edu (Evan Hanusa), deepak@ee.washington.edu (Deepak Vijaywargi), bilmes@ee.washington.edu (Jeff Bilmes), fox@cs.washington.edu (Dieter Fox)

on audio with noise conditions that are matched to the audio being recognized is likely to be an upper bound on the performance of model compensation schemes (acoustic model adaptation) [1]. The use of training audio that exhibits the Lombard effect has also been shown to improve the performance of speech recognition systems [7].

Many speech corpora have been developed for studying algorithmic improvements that provide ASR performance increases; however, few of them provide ideal training data for systems that must recognize conversational speech in noisy environments. For example, TIMIT [8] consists of read speech with no noise, and Broadcast News [9] contains mostly read speech. Limited-vocabulary corpora exist, such as NOIZEUS [10] and the AURORA databases [11], which contain recordings of spoken digits (and other small-to medium-vocabulary settings) that are clean, recorded in noisy environments, and/or often artificially distorted (by additive noise and simulations of rooms and telephone networks), as well as SPINE [12], in which background noise was played on a speaker in the recording booth. Other corpora have been designed to capture the Lombard effect, including UT-Scope [13] and the Albayzin Spanish-language corpus [14]. The ICSI [15] and AMI [16] meeting corpora contain microphone array recordings of multi-party conversations in indoor environments. Several in-car corpora have been created, with multi-microphone recordings of limited-vocabulary speech in noisy environments. These include AVICAR [17] and the CIAIR Japanese corpus [18], which also includes dialog recordings. There are also databases which capture the effects of specific types of distortion, such as telephone channels in Switchboard [19]. Additionally, some multi-modal corpora exist (including AVICAR and the AMI corpus), that allow the combination of, say, audio and visual information.

Our goal was to create a corpus that brings together many of the elements that make each of these corpora useful: i.e., the presence of various levels and types of background noise, recordings of Lombard speech with and without the background noise, spontaneous multi-person conversations, and synchronized multi-microphone recordings (including a microphone array) of each conversation participant. Many considerations motivated the design of the recording hardware and data collection practices. The corpus contains multi-party conversations about everyday topics. Additionally, the speech is recorded in true noisy environments, rather than having the noise added later or piped in as background noise through speakers. In fact, the participants are actually walking around *within* the real noisy environment in which the speech is recorded, and are potentially being affected by all of the distractions that such a noisy-environment might entail. These noisy environments range in both noise type and intensity, and include a wide range of indoor and outdoor noise sources such as crowds, vehicles, and wind at a variety of SNRs. To achieve this, we have custom-designed a portable recording device that allows for the multi-track speech to be recorded *in-situ*, rather than making the recordings in a studio, which would affect the speakers' comfort, behavior, and speech patterns. Of course, any speech that compromises the privacy of the speakers is deleted from the corpus, but the fact that there is such speech indicates that the speakers engage in comfortable, natural, and highly disfluent speaking styles, conversations, and vocalic patterns. We therefore believe that our corpus, due to its inherent in-situ nature, provides a unique perspective on aspects of human speech production when spoken in real-world noisy environments, and also on the acoustic properties of speech and noise when it is collected in such real-world environments.

Our resulting "COSINE" corpus was first introduced in [20]; the present paper gives a much more detailed description of every aspect of the creation of the corpus. The paper is organized as follows: In Section 2, we discuss the design and construction of our custom portable wearable recorders. In Section 3, we describe the recording sessions during which the corpus was recorded. Section 4 explains the word-interior annotation method that was used to mark words during transcription, and the nature of the transcription is described in detail in Section 5. Section 6 contains details about the public release of the corpus. In Section 7, we conclude with a discussion of potential applications for the resulting recordings.

2. Portable wearable multi-channel recording system

We designed the hardware component of our speech recording system to be portable, light-weight, unobtrusive, and comfortable. Given the requirements of the corpus mentioned in the previous section, this

involved evaluating different trade-offs between microphone placement and audio quality. In order for the audio in the corpus to represent a variety of scenarios, we chose to record the speech with several different types of microphones, which cover a broad range of comfort, audio quality, and noise rejection. To capture multi-person conversations in a natural environment, we needed a portable system capable of synchronously recording many channels. However, most multichannel recording systems are not designed for portability, and commercially available recording systems that are portable and have more than two microphone inputs are quite costly. As a result, we designed a custom wearable recording setup with seven microphones for our data collection.

Our recording system consists of a lightweight backpack, an array of four microphones positioned in front of the speaker’s chest and directed at the speaker’s mouth, a close-talking microphone, a throat microphone, an electret microphone mounted on the shoulder strap, and two modified Zoom H2 four-channel, 24 bit, 48 kHz audio recorders (see Figure 1). It is critical to realize that we designed our device to collect speech using a large diversity of different recording modalities. This was done to maximize the amount of potential speech research that could benefit from our corpus. We do not, of course, expect our wearable device to be a prototype for a final portable electronic recording device, but it might be a reasonable surrogate for a super-set of such a prototype, and we moreover do expect that the resulting research made possible from this corpus could benefit such a device.

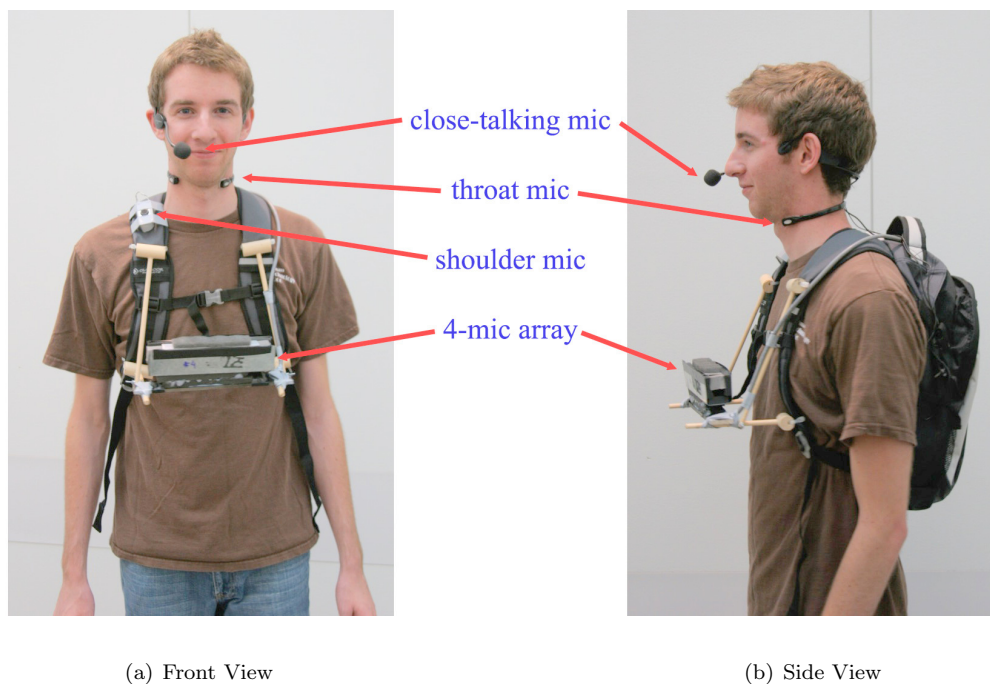


Figure 1: The wearable recording system

A broad range of audio quality is covered by the various microphones in the system. The Sennheiser ME-3 close-talking microphone is the standard for desktop speech recognition, and provides very high quality speech recordings, with good rejection of background sounds. The Fire Fox Sportsman throat microphone records a distorted version of wearer’s speech, yet has near-ideal background noise rejection. Depending on microphone placement and the wearer’s physiology, the distortion can be minimal or can cause near unintelligibility. The electret microphones on the shoulder and within the array also record high quality audio, but allow a substantial amount of background noise, especially of neighboring speakers. However, they are very compact (3/8” diameter), and are perhaps the most practical type of microphones to use in a system that presents no comfort burden to the wearer. The array configuration allows the use of post-processing to focus the array’s beam onto or away from the speaker wearing the device, and could

conceivably be built into a compact, wearable system.

2.1. Four-microphone array

The system contains an array of four microphones, worn in front of the speaker’s chest. Array processing would be impossible without extremely precise clock-level channel synchronization. Since the maximum number of channels available on any single-clock portable recording device was four, this restricted the size of the array, also to four. The design and construction of the microphone array are detailed in the following sections.

2.1.1. Microphone selection

Several microphone types were considered for use in the four-microphone array. Background noise rejection is important, so unidirectional (rather than omnidirectional) mics were evaluated. In addition to testing the microphones from the Zoom H2 recorder which was selected for our system, we purchased three models of unidirectional electret microphones, and measured their directionality by playing back white noise in an anechoic chamber, varying the angle of incidence of the sound to the microphone, and comparing the energy of the resulting recordings. The measured response of the various microphones is shown in Figure 2. Glancing angle is 90° minus the angle of incidence. High attenuation at low glancing angles is desired, and the sensitivity of the stock H2 microphones is within 0.3 dB of the best microphone in this regard, the DUM5246. Since the impedances of the microphone models differ, we chose to use the stock microphones to avoid any mismatch issues between the microphones and the recorder. The beam pattern (normalized directional amplitude response) of the selected microphone (for white noise) is shown in Figure 3(a).

2.1.2. Microphone spacing

The characteristics of a beam formed by a microphone array depend on microphone spacing and the frequency of the sound. For a fixed spacing, the beam is widest for low frequencies and is narrower for higher frequencies. Choosing microphone spacing involves a design tradeoff: wider spacing results in a higher-resolution array (with a narrower beam over the entire frequency range), but lowers the frequency above which spatial aliasing (the presence of significant sidelobes at high frequencies) occurs. The overall size of an array that must be comfortably wearable also restricts the maximum microphone spacing.

To completely avoid spatial aliasing, the distance between microphones should be $d < \lambda_{min}/2$ [21]. For speech processing, it is desirable to capture frequencies up to 8000 Hz, which corresponds to a minimum wavelength $\lambda_{min} = 4.25$ cm, or a microphone spacing of $d < 2.125$ cm. However, [22] found that making the microphone spacing too small or too large degraded speech recognition performance. In their experiments, cross-correlation processing and delay-and-sum beamforming were used to combine channels from a 7-microphone linear array, and a spacing of 6-8 cm gave the best word recognition accuracy for both methods.

For our 4-microphone system, a microphone spacing of 3 cm was chosen as a good balance between resolution, spatial aliasing, and overall array size and comfort. This spacing results in an array span of 9 cm, and the total width of our array, including the housing, is 20 cm; any wider would be unwieldy and would interfere with natural arm movement when worn.

The beam pattern of an ideal array of four omnidirectional microphones with 3 cm spacing is shown in Figure 3(b). The beam pattern of such an array, adjusted for the measured directionality of the H2 microphone, and aimed at 90° (straight ahead), is shown in Figure 3(c). Due to the directionality of the microphones, the shape of this beam pattern would change when the beam is aimed in any direction other than 90° .

2.1.3. Microphone array materials and construction

The goal of the recording system is to capture the speaker as well as the background sounds and to completely reject any noise due to two main sources: tapping or touching the cable that connects the array to the recorder and bumping the array housing. These noise sources are common because individuals wearing the system are expected to be moving during their conversations. Many iterations of the array design were required to arrive at a design with acceptable rejection of these types of noise.

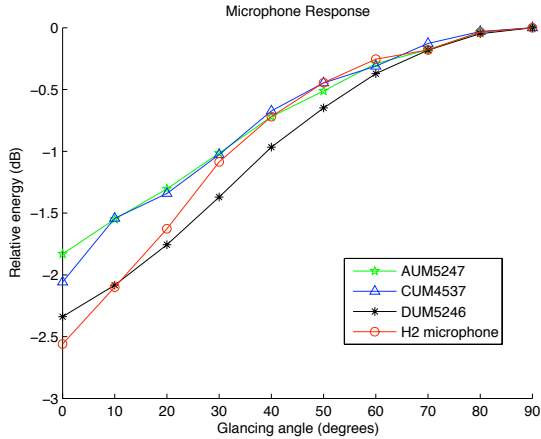


Figure 2: Directionality of microphones considered for the microphone array.

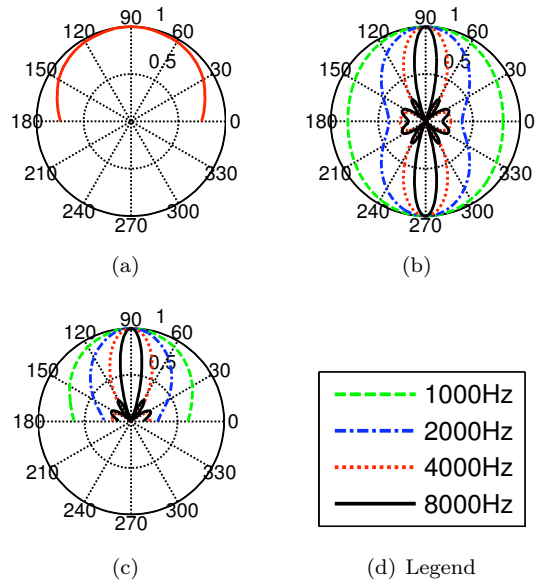


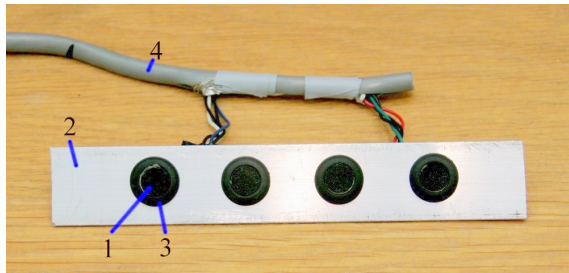
Figure 3: Beam patterns of: (a) one H2 mic (for white noise), (b) An array of four ideal omnidirectional mics with 3 cm spacing (for four frequencies), and (c) an array of four H2 microphones with 3 cm spacing (for four frequencies).

In our final design, shown and labeled in the four photos in Figure 4, the microphones (1) are placed within holes drilled in a 6" x 1" strip of aluminium (2). To acoustically isolate the microphones, and to precisely position them within the array, they are placed into rubber grommets (3) which line the holes. The grommets hold the microphones snugly in place and help to insulate them against vibration.

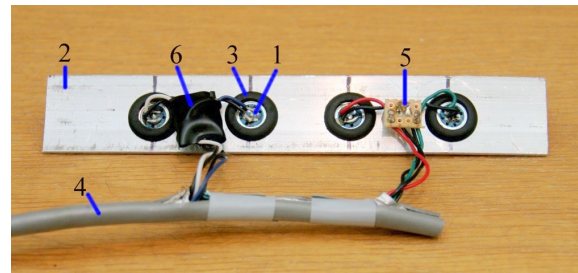
Several methods were attempted to reduce the effect of sound waves traveling along the cable (4). Initially, we tested different combinations of gluing and clamping the cable or the wires within it to a piece of lead. A more effective solution was used for the final design: thin wires are attached to the microphones and soldered to a small (1 cm x 1 cm) PCB (5), which is then soldered to the wires within the cable (4) that connects to the recorder, and wrapped with electrical tape (6). The PCB is compact and lightweight, and its purpose is to absorb most of the vibrations traveling up the cable, keeping them from reaching the microphones. The PCB assembly is shown between each pair of microphones in Figure 4(b).

The second type of undesirable noise can come from the microphone array housing or its supports being touched. The proper choice of materials was necessary to minimize vibrations conducted through the array housing (7,8,9) and then picked up by the microphones. Initial designs used balsa wood and plastic as the structural material, since they are lightweight and easy to shape. However, these low density materials are good conductors of sound, and any rubbing or tapping of the structural material was very audible in the audio recordings. Lead was also tested, since it is very dense, but it was challenging to securely attach the microphones and wires to a piece of lead in a way that would absorb vibrations. Lead was also found to be prohibitively heavy.

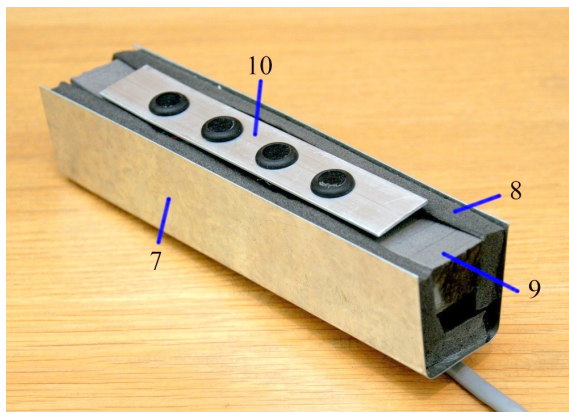
Acoustic foam was found to be a good material but has the disadvantage of being flexible, so a structural support was necessary. Several types of foam were tested, and various configurations of plastic and aluminum were initially tried as supports. The final design uses a thin U-shaped steel plate (7) as the outer structural support (shown in Figure 4(c)), lined with 3/8" adhesive rubber foam seal (8), and with a 3/4" x 1" strip of EVA foam (9) held in place in the middle of the assembly. The two types of different density foam act as sound-deadening materials. The cable is passed between the two types of foam, and the aluminium strip/rubber grommet/microphone assembly (10) is glued down on top of the EVA foam using hot glue, sandwiching the tape-wrapped PCBs between the EVA foam and the aluminium. Finally, a wind screen of



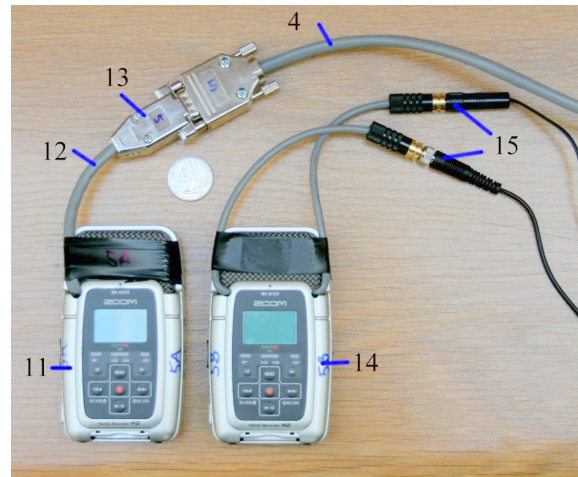
(a) Top view



(b) Bottom view



(c) Array assembly



(d) Zoom H2 recorders

Figure 4: Parts of the microphone array and the recording system. (1) electret microphone, (2) aluminium strip, (3) rubber grommet, (4) 3 ft shielded audio cable, (5) PCB, (6) electrical tape, (7) U-shaped steel plate, (8) adhesive rubber foam seal, (9) EVA foam, (10) aluminium strip/rubber grommet/microphone assembly, (11) array mic recorder, (12) short cable, (13) DB-9 connector, (14) close/shoulder/throat mic recorder, (15) mini-jack connectors.

3/8" low-density foam (not shown) is affixed to the top of the aluminum to reduce wind and breath noises.

A minimal length of wire between the shielded cable and the PCBs, as well as the PCBs themselves do not have shielding; however the steel and aluminium components of the enclosure (Figure 4(c)) provide some shielding. Sample recordings on the resulting devices showed that shielding from electromagnetic interference is adequate and that there is no crosstalk among the individual channels in each recorder.

2.2. Selection of recorders

The following criteria were considered when choosing recording devices: number of channels, portability, battery life, recording quality (16/24 bit), sampling rate, compatibility with close-talking and throat microphones, amount of modification necessary, cost, and synchronization issues.

There are many relatively inexpensive portable two-channel recorders on the market. However, all channels of the microphone array must be recorded either on one device, or on multiple recorders with perfect synchronization. Recorders that can run off an external clock are prohibitively expensive (\$999 for the cheapest 2-channel device available in 2008), so we were limited to the following options: 1) a laptop and a USB/firewire audio interface, 2) a custom-built recording solution, or 3) a modified four-channel portable audio recorder.

The cheapest semi-portable recorder with 4 external microphone inputs on the market in 2008 was the Edirol R-4 (\$999), and options 1 and 2 are also prohibitively expensive. The Zoom H2 recorder (\$199), after some modifications, proved to be a good solution. This recorder is compact, battery powered, can record

at 24 bit/48 kHz, and has four internal electret mics, each recorded on its own channel. Because there is no four-channel input jack on the device, the recorders were opened, the microphones were unsoldered, and connectors for external microphones were soldered in their place.

In each recording setup, one of the two Zoom H2 recorders (11) is configured to capture the audio from the microphone array. The four unidirectional electret microphones are taken from inside the recorder, and mounted in the external microphone array. The wires of a short shielded audio cable (12) are soldered to the microphone terminals in the recorder, and a DB-9 connector (13) is attached to the other end of the cable. The cable is secured inside and outside of the recorder so that cable movement does not damage the solder connections. A 3 ft length of shielded audio cable (4) with a twisted pair for each channel is used to connect the array to the recorder via the DB-9 connector (Figure 4(d)).

The second H2 recorder (14) captures the audio of the close-talking microphone, the throat microphone, and the shoulder-mounted electret microphone, which is taken from inside the recorder. These three microphones are connected to the H2 recorder using shielded wire with secure screw-on mini-jack connectors (15). The shoulder microphone is mounted in a rubber grommet within a hole in a 2" x 1" piece of aluminium, which is hot-glued to a piece of EVA foam, and the entire assembly is secured by electrical tape to the shoulder strap of the backpack (see Figure 1(a)). Since the Sennheiser ME-3 microphone requires power to operate, the "+" terminal from the fourth audio channel is used to provide power to the ME-3, so only three channels are available for recording audio on this device. Along with the four channels on the other device, this gives seven channels for each speaker.

The devices are configured to record in 24 bit/48 kHz mode, and recorder gain is set to the lowest setting so that minimal clipping occurs, even at high volume levels. The recorders store the 4-channel data on 4GB SD cards. This results in approximately 2 hours of continuous 4-channel recording. By recording at 24-bit resolution, the low gain setting is not detrimental to the quality of the recording. After recording, the audio can be amplified, limited (soft or hard), and then converted to 16-bit resolution. For the released version of the corpus, the bit rate is reduced to 16 bits, with rescaling (the close and throat microphone recordings are amplified by 9 dB; the others are not amplified), soft limiting, and dithering prior to bit depth reduction, and the sampling rate is reduced to 44.1 kHz.

2.3. Synchronization

When making recordings on multiple devices, it is important to synchronize the audio streams from the two recorders worn by one person, as well as all the recorders worn by all participants of one session. For each speaker, all of the array channels are recorded on one device, achieving sample-level synchronization. The other three channels are recorded on a second device, so synchronization between these three channels is also ideal. Two factors must be accounted for in synchronizing devices: the start time, and any sampling rate differences between devices. Even a small sample rate variation between devices can lead to several seconds of desynchronization after several hours of recording, and we found this to be the case. Inter-device synchronization can be achieved using synchronization tones, played at the beginning and end of the recordings. A matched filter can be used to detect the tones, and then each device's audio can be resampled if necessary. This method was used to synchronize the recordings in our corpus. Tones were played at the beginning and end each session, and were located to within 3 ms. Tones at the end of each session were not successfully recorded for a few of the sessions, due to technical errors, so other sounds were used to find a common reference time near the end of each session in these cases. These reference times are accurate to within 30 ms. Prior to resampling, we measured the difference between the length (between the start and end tones) of each audio recording in the corpus and the length of the shortest recording in that session. This difference was normalized by the length of the session, and the normalized difference (in seconds per hour) was found to have a mean of 0.074, a variance of 0.013, and a maximum of 0.607. The located synchronized start and end points were used to align the beginnings of the sessions and adjust the sample rates so that all audio became synchronized throughout. We also verified that the sample rate of any particular recorder does not vary significantly over the course of a two-hour session, so aligning the start and end tones is sufficient for ensuring proper synchronization throughout the entire duration of the recordings.

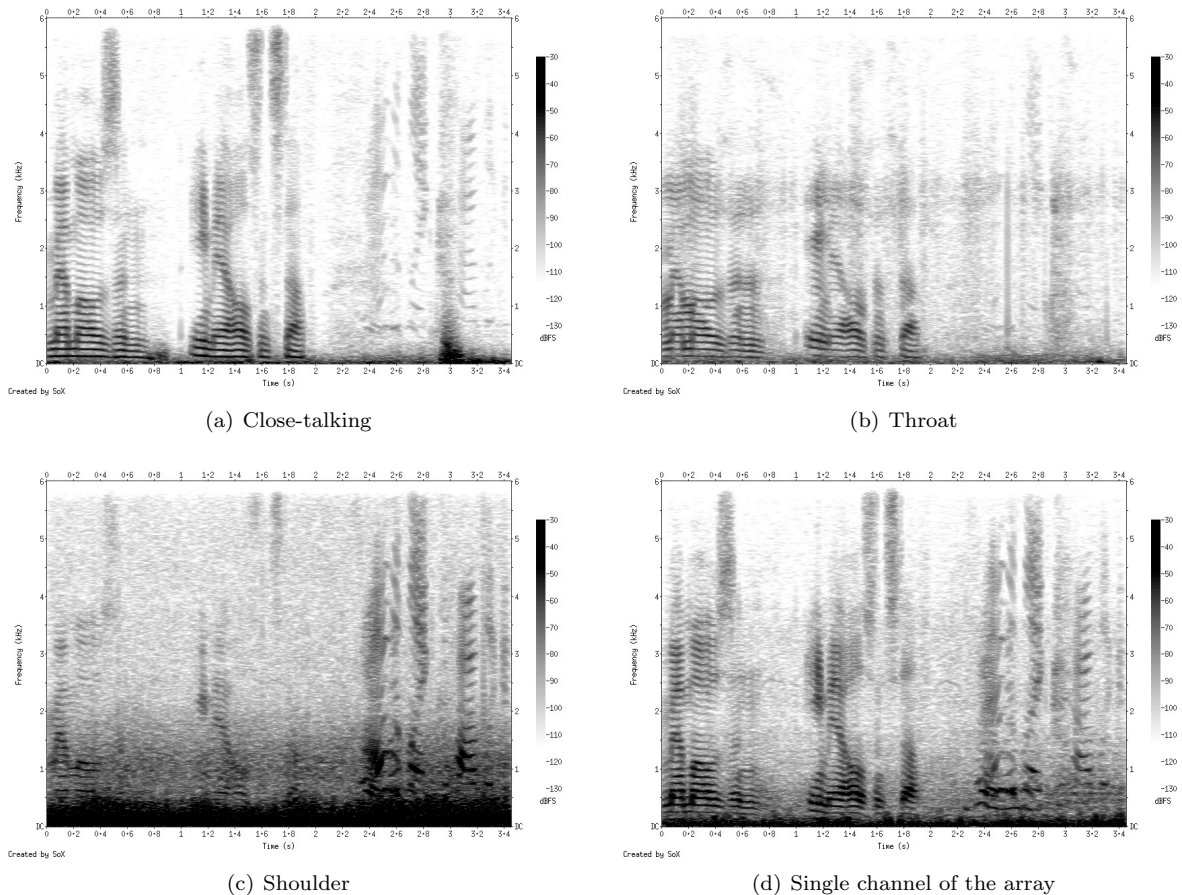


Figure 5: Spectrograms of a two-person dialog. The person wearing the microphone says “memos and emails and stuff” (the first two seconds), followed by a bystander saying “I would like to have like”. Note the relative level of bystander speech in each microphone.

2.4. Properties of audio recordings

The varying positions and types of microphones result in different audio characteristics in the separate recording channels. The spectrograms of Figure 5 show a recording with speech of the person wearing the device (first two seconds) and another person standing several feet away (the remainder of the duration). The characteristics of the throat microphone are immediately noticeable: it completely rejects the background noise and the sound of the bystander in the second half of the recording, and there is very little energy content above 4 kHz. The close-talking Sennheiser microphone picks up only a small amount of the background noise, and only small portions of the bystander’s speech. The array microphone picks up more background noise, as well as the bystander speaking. Audio quality from the shoulder microphone can vary greatly depending on the environment - it lacks a wind screen, and in this situation it picks up lots of low-frequency wind noise.

3. Recording sessions

Paid volunteers (approved for human subjects research) participated in multi-person recording sessions that lasted between 45 minutes and 1.5 hours. The breakdown of the number of people per session is: 2 people: 13%, 3 people: 19%, 4: 42%, 5: 3.5%, 6: 19%, 7: 3.5%. Even numbers of participants were favored because people tend to talk more when they are able to pair up. After putting on the recording

devices, the volunteers were asked to walk to various noisy locations, and to talk about anything they like. Both pairwise conversations and group discussions were encouraged. The participants were given a list of suggested open-ended conversation topics to use in case they ran out of things to talk about, though this was rarely necessary. As a result, the conversations are spontaneous, colloquial, and natural.

A total of 33 sessions were held, 10 of which were transcribed (see Section 5 below). The transcribed sessions lasted a total of 11.40 hours, yielding 42.10 hours of 7-channel audio, with an average of 3.69 participants per session. The untranscribed sessions lasted a total of 26.27 hours, yielding 110.86 hours of audio, with an average of 4.22 participants per session. The lengths of audio and various other statistics of the recordings are shown in Table 1.

3.1. Speakers

All the speakers whose voices were recorded are fluent (but not necessarily native) English speakers. There are 91 unique speakers, 59 of whom participated in one session, 22 participated in two sessions, and 10 participated in 3 sessions. The transcribed sessions have 37 unique participants, each of whom participated in only one transcribed session. The speakers’ ages range from 18 to 71, with a median of 21 and a mean of 25. Each speaker filled out a survey about their experience learning and speaking English, and the answers are included along with the audio recordings and transcripts. The number of hours recorded by speakers of each gender is shown in Table 1.

3.2. Noise types

The recordings were made indoors and outdoors, on and near the campus of the University of Washington in Seattle, WA, during daytime hours between July and September of 2008. The subjects walked around during the sessions, which affected the nature of their speech and the recordings in general. They were free to walk anywhere, but were instructed to spend as much time as possible in environments with significant amounts of background noise. Typical destinations included streets and public areas on the university’s campus, as well as the main commercial thoroughfare near campus, where most ground floor spaces are shops and restaurants, and which is heavily trafficked by pedestrians, cars, and buses. Many types of background noises are represented, including: bus and car engine sounds while walking along streets, noise from construction sites, water from a large fountain, birds, wind noise, and people in a busy cafeteria at lunchtime.

As a consequence of the large amount of movement the recorders had to withstand, the wiring for the throat and shoulder microphones in some recorders developed a short or poor contact, causing audio to cut out. Fortunately, such problem regions are relatively infrequent, and are marked in the annotation files that accompany the audio. The durations of affected audio from each microphone are shown in Table 1. These problems did not occur with the array microphones, and happened in very few close-talking microphone recordings.

		Male	Female	Total	Mic problems		
					Close	Throat	Shoulder
Transcribed sessions	Speech	4.55	7.80	12.35	0.02	0.88	2.09
	Nonspeech	12.27	17.48	29.75			
	Total audio	16.82	25.28	42.10			
Untranscribed sessions	Total audio	43.78	67.08	110.86	1.35	3.97	7.47

Table 1: Corpus audio statistics (in hours)

4. Word-interior annotation

An important consideration for the corpus was the extent to which the data would be labeled. To expedite the transcription process, three methods were evaluated:

1. fully-labeled (FL) - transcribers mark the precise beginnings and ends of words,
2. sequence-labeled (SL) - transcribers mark the beginning and end of a phrase and then transcribe only the sequence of words, and
3. a technique introduced in [23] called partially-labeled (PL) - the word sequence is transcribed, and an identifying mark is placed somewhere within each word.

As shown in [24], the PL method of annotation is significantly faster than the FL method (0.052 words/second for FL vs 0.134 words/second for PL), and results in improved WER performance over both when training and testing on Switchboard conversational telephone speech data. While recent results have shown that the PL approach can be approximated by using a two-pass training strategy [25], PL can benefit from human annotators since they are immune to out-of-vocabulary lexical items and disfluencies that become more frequent as the colloquial nature of the speech in the corpus, such as our own, and any noise in the corpus, increases. In our annotation process, transcribers reported only a 40% speedup when annotating using the SL method compared to the PL annotation method. Because the PL data results in improved WER over the SL and FL data, the PL method was used for the COSINE corpus. Figure 6 shows an example of a PL annotation with a privacy deletion (discussed in Section 5.3).

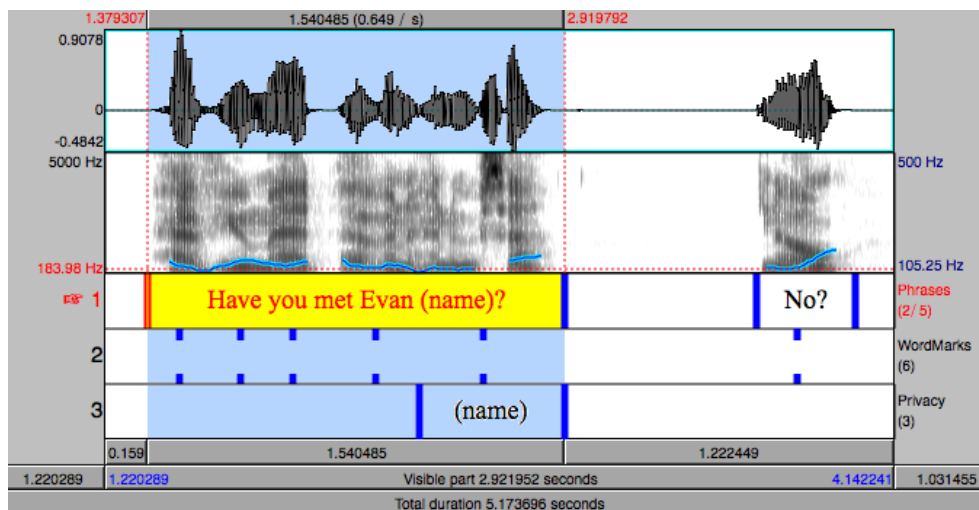


Figure 6: Example annotation in the Praat program. Tiers 1, 2, and 3 show phrases, PL word marks, and privacy regions, respectively.

5. Orthographic transcription

Using the Praat program [26], a three-pass transcription of the corpus was done by a group of twelve 3rd and 4th year Linguistics undergraduate students at the University of Washington. All transcribers are native English speakers.

To ensure consistency of transcription, an annotation guide was created for the transcribers to follow. A wiki was also established, allowing the transcribers to standardize their transcriptions and to learn from each other. The wiki contained discussions of how to properly transcribe non-obvious cases, as well as constantly growing lists of slang; proper nouns; neologisms; non-standard pronunciations of standard words; shortenings or compounds that were not found in the dictionary; and compound words or expressions whose pronunciations differ from the concatenation of the pronunciations of their constituents. The resulting list has over 500 terms, and is included in the corpus release. The wiki was frequently updated, and served as a very useful tool throughout the entire transcription process.

Prior to beginning work on the corpus, the transcribers spent several weeks transcribing other speech recordings. During this time, the wiki and the guide underwent revision as rules for transcribing various

significant phonetic and phonological phenomena were standardized. Transcription of the corpus began only after the transcription rules had stabilized and all transcribers were familiar with the process.

A measure of inter-annotator agreement was calculated as follows. Each transcriber annotated the same half-hour section of conversational speech. The text was stripped of all punctuation and other symbols described in the transcription guide (Section 5.2). A modified edit distance between each pair of transcriptions was calculated. It used the normalized character edit distance between words as the substitution cost in order to reduce the cost of errors such as substituting “hah” with “ha”. The edit distance between two files was normalized by the maximum possible edit distance, and the resulting average annotator disagreement was 9.2%. While this number may appear high, most of the errors were relatively minor due to our not performing ideal text normalization before measuring the modified edit distance - we therefore take this number as an upper bound on the true inter-annotator error rate. Also, this measurement was performed only after the first pass of transcription; the two additional passes (see Section 5.1 for a description of the 3-pass process) are likely to have further increased the consistency between the transcriptions of different recordings.

5.1. Transcription process

The transcriptions are intended to satisfy two criteria: first, they must properly identify the word that was said (even if it was pronounced unconventionally), and second, they should indicate whether or not a word is pronounced correctly (see Section 5.2 for detailed explanation). Many of the transcription features are similar to the AMI Corpus transcriptions [16].

During the first pass of transcription, transcribers listened to 15 minute segments of conversations from the close-talking microphones and marked the boundaries of intervals corresponding to segments of speech, transcribed the speech in each interval, and placed a mark in the interior of every word in the transcript. Segments of silence shorter than 1/2 second that occurred between words or sentences were not required to be marked as separate silence regions. Before the submission of a completed annotation segment, a spellcheck ensured that all words are in the cmudict0.7a dictionary [27] or one of the word lists in the wiki (referred to as the “dictionary appendix”), and the transcription was checked to ensure that the number of word interior marks in each speech interval is correct. Transcription rate was measured to be approximately 20 hours of work to transcribe 1 hour of speech.

A second pass was performed over all transcriptions by four of the transcribers, each transcriber working on files that they did not initially transcribe. During the second pass, transcribers listened to all the audio and corrected any mistakes they encountered. A final third pass was performed to further improve the quality and consistency of the annotations, catching any remaining mistakes. For the second and third passes, all areas of the transcription that violated any transcription syntax rules or had misspelled words (words absent from the dictionary or appendix) were automatically highlighted, to bring the transcribers’ attention to these potentially problematic areas.

5.2. Transcription guide

The words in each region are transcribed in standard American English. Transcribers had access to the CMU dictionary (cmudict0.7a), and were able to search through it to determine the presence or absence of a word. They also made use of dictionary.com to verify spelling. Punctuation (. , ; ? !) is included in the transcriptions, and standard capitalization rules were followed.

The following is the process flow that the transcribers used to annotate the data. First, the transcriber would listen to a phrase, transcribing the phrase at the word sequence level, and creating a word interior mark for each word (Figure 6). Any word that is completely unintelligible is marked with a “+”. A sequence of unintelligible words is marked with repeated “+” symbols, and each “+” is given a single word interior mark.

Once the phrase has been transcribed, words that are “mispronounced” are appended with a “*” (an asterisk). Whether or not a word is “mispronounced” is largely subjective, and is based on individual transcriber opinion. In some cases, it is possible to understand the sentence that is spoken in the recording, but not necessarily understand individual words. To account for this, any word that is not identifiable when

listened to in the context of one preceding and one following word is marked with a “*”. In cases where a speaker uses a pronunciation of a word that is uncommon and not present in the CMU dictionary (which often has multiple pronunciations for one word), the transcriber is instructed to do the following: if the speaker always mispronounces the word in the same way, the word is not marked with a “*”, as it may be a dialectal difference. If the speaker does not consistently pronounce the word in this manner, the word is marked with a “*”. Lastly, if the word is pronounced in a common fashion, or if its pronunciation can be derived from the dictionary pronunciation through allophonic variation or standard phonological rules (such as “T” sounds like “D” in “little”, or “D” sounds like “J” in “did you”), it is not marked with a “*”.

The asterisks are included in the transcriptions for two reasons. Severely mispronounced words would negatively impact any acoustic models trained on the data. Words marked with a “*” can either be ignored for this purpose or can be treated in a special way, as the researcher desires. Mispronounced words are not excluded completely because the transcripts can be useful for data-driven pronunciation model research as well as training of language models or other text-based research, also important in speech recognition [28].

Acronyms are indicated by capitalization. Any word in the transcript that is fully capitalized is an acronym pronounced as the sequence of letters (FBI, NBC, GPS, etc.). Acronyms that are not pronounced in this manner are indicated by a fully capitalized word followed by a “~” (a tilde), and are added to the dictionary appendix. Some examples are “NAFTA~” and “NASA~”. Acronyms with and without a ~ can be made plural or possessive, such as “MCATs~” or “FBI’s”.

Additional special symbols are introduced: “\$” is used to indicate laughing. It can be a standalone symbol or can be appended to a word to indicate laughing and talking simultaneously, such as “really\$?”. “#” is used to indicate whistling, coughing, sneezing, non-verbal singing, or miscellaneous vocal noise. “@” is used to indicate a foreign word. It can replace an unknown foreign word or be appended to a word (e.g., “bonjour@”). It is *not* used if the word is pronounced using American English phonemes and phonological rules. As a result, foreign words that are part of the standard American English lexicon, e.g. “burrito,” are not appended with a “@” when spoken with an American English pronunciation. “-” is used to indicate a disfluency or discontinuity at the beginning or end of a word. When a word is not wholly pronounced, the pronounced part is typed out, followed by a “-” (e.g., “I think basi-”). All names of people, places, organizations, and events (i.e., named entities) are preceded with a “^”. Common nouns acting as a name of person, place, etc. are also marked with a “^”. If the name contains more than one word, all words in the name (including function words) are annotated with a “^”.

Singing or melodic speaking is not annotated in any special way. Commonly occurring suffixes (’ve, ’ll, ’s) are transcribed as part of the word, but are excluded from the word when performing a spellcheck. Other less common suffixes (such as “y”) may be separated from the root, if the root appears in the dictionary or appendix, but the entire word does not, e.g. “orange ’y type dressing.”

5.3. Privacy

To protect the privacy of the subjects, all occurrences of privacy-sensitive speech in the recordings have been deleted – the audio signal in the recordings of all conversation participants during these times is set to zero, and in the transcripts, all words in these phrases are replaced by a privacy token which describes the deletion. These tokens are: “(name)”, “(place)”, “(phone)”, “(number)”, and “(other)”. Information that is considered to be private is: last names (except of public figures), addresses, phone numbers, account numbers or PINs, anything else that may reveal a speaker’s identity, and admissions of illegal activity. In the case of illegal activity, a minimal amount of speech is removed to protect the speaker, for example “I saw John (name) (other) that car”. Also, any information that the subjects explicitly asked to delete, regardless of its nature, was removed. The need for this type of deletion is an indicator of the type of real-world conversations that are captured in the COSINE corpus. Subjects are at ease because they are not constrained to a studio environment. Due to the nature of the corpus, a privacy-related deletion will span the recordings of all the subjects who were potentially in a conversation with the person who divulges private information. There are 144 seconds of privacy deletions in the transcribed audio, for an average of 3.41 seconds in every hour of audio. The untranscribed sessions contain a total of 1136 seconds of privacy deletions, for an average of 10.25 seconds per hour of audio.

6. Release

The final release of the corpus contains all of the recorded audio (excepting any privacy-related deletions), at a 44.1 kHz sampling rate and a bit depth of 16 bits, and stored in the FLAC compressed lossless audio format [29]; the transcriptions; and all non-privacy-sensitive subject information. The corpus is available online¹ to speech researchers free of charge.

7. Conclusion

We expect that the COSINE corpus could be a unique and valuable tool for the speech and language community. Its annotations comprise word-level transcriptions of multi-party in-situ conversational speech, including word-interior markings. Each speaker has been recorded simultaneously on seven different channels with noise content and channel distortion, representing the varying conditions of real-world microphone types and placement. As the speech has been recorded in-situ, there are no artifacts from adding noise to conversations in a studio environment or after the speech is collected; these conditions are not broadly available in other corpora. The multi-party conversations in the corpus are unprompted, resulting in spontaneous and natural conversation.

The COSINE corpus can be used for a variety of speech-related research, such as improving techniques for noise reduction in audio, and feature cleaning/combination/transformation/design. The array audio can be processed with classical beamforming, adaptive beam steering, or other techniques such as Likelihood Maximizing Beamforming [21], as well as new methods. Schemes to map directly between features from the noisy throat or far-field microphones, and features from the clean, high quality close-talking microphone, can be developed, for speech denoising using regression models [30]. Techniques for combining data streams at various stages of a speech recognizer (the audio level, feature level, and recognizer output) can also be investigated. The synchronization of the audio from multiple speakers also allows this data to be used in research on dialog acts and conversational dynamics, such as described in [31].

8. Acknowledgements

This material is based upon work supported in part by DARPA’s ASSIST Program (contract number NBCH-C-05-0137) and an ONR MURI grant (No. N000140510388). The authors would like to express their thanks to the transcribers who made this work possible: Naomi Bancroft, Eric Braun, Dutch Hixenbaugh, Alexander Keane, Brent Nelson, Kellen Michael Paisley, Justina Rompogren, and Yeon-Hee Yim, with special thanks to the four transcribers who also did additional quality screening of the data: Min Amodio, Alana Katchuk, Cari McLean, and T.J. Trimble.

References

- [1] Y. Gong, Speech recognition in noisy environments: a survey, *Speech Communication* 16 (3) (1995) 261–291.
- [2] W. Guo, L. Zhang, B. Xia, An auditory neural feature extraction method for robust speech recognition, in: *ICASSP*, 2007.
- [3] C. Chen, Noise robustness in automatic speech recognition, Ph.D. thesis, University of Washington (2004).
- [4] M. Islam, H. Matsumoto, K. Yamamoto, An improved mel-Wiener filter for mel-LPC based speech recognition, in: *Interspeech-ICSLP*, 2006.
- [5] W. Lim, C. Han, J. Shin, N. Kim, Cepstral domain feature compensation based on diagonal approximation, in: *ICASSP*, 2008.
- [6] D. Yu, L. Deng, J. Droppo, J. Wu, Y. Gong, A. Acero, A minimum-mean-square-error noise reduction algorithm on mel-frequency cepstra for robust speech recognition, in: *ICASSP*, 2008.
- [7] R. Lippmann, E. Martin, D. Paul, Multi-style training for robust isolated-word speech recognition, in: *ICASSP*, Vol. 12, 1987, pp. 705–708.
- [8] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, N. L. Dahlgren, *DARPA TIMIT Acoustic Phonetic Continuous Speech Corpus CDROM* (1993).

¹<http://ssli.ee.washington.edu/cosine> or web-search for “cosine speech corpus” to find the corpus’s current residence.

- [9] D. Graff, Z. Wu, R. MacIntyre, M. Liberman, The 1996 broadcast news speech and language-model corpus, in: Proceedings of the DARPA Workshop on Spoken Language technology, 1997, pp. 11–14.
- [10] Y. Hu, P. Loizou, Subjective evaluation and comparison of speech enhancement algorithms, *Speech Communication* (49) (2007) 588–601.
- [11] Aurora speech recognition experimental framework, <http://aurora.hsnr.de/>.
- [12] A. Schmidt-Nielsen, T. Crystal, E. Marsh, Speech in Noisy Environments (SPINE) adds new dimension to speech recognition R&D, in: HLT, 2002.
- [13] V. Varadarajan, J. Hansen, I. Ayako, UT-SCOPE—a corpus for speech under cognitive/physical task stress and emotion, in: The Workshop Programme Corpora for Research on Emotion and Affect, 2006.
- [14] A. Moreno, D. Poch, A. Bonafonte, E. Lleida, J. Llisterra, J. Marino, C. Nadeu, Albayzin speech database: Design of the phonetic corpus, in: EUROSPEECH, 1993, pp. 175–178.
- [15] A. Janin, D. Baron, J. Edwards, D. Ellis, D. Gelbart, N. Morgan, B. Peskin, T. Pfau, E. Shriberg, A. Stolcke, C. Wooters, The ICSI Meeting Corpus, in: ICASSP, Vol. 1, 2003, pp. 364–367.
- [16] J. Carletta, S. Ashby, S. Bourban, M. Flynn, et al., The AMI meeting corpus: a pre-announcement, *Lecture notes in computer science* 3869 (2006) 28.
- [17] B. Lee, M. Hasegawa-Johnson, C. Goudeseune, S. Kamdar, S. Borys, M. Liu, T. Huang, AVICAR: audio-visual speech corpus in a car environment, in: ICSLP, 2004.
- [18] N. Kawaguchi, S. Matsubara, H. Iwa, S. Kajita, K. Takeda, F. Itakura, Y. Inagaki, Construction of speech corpus in moving car environment, in: ICSLP, 2000.
- [19] J. Godfrey, E. Holliman, J. McDaniel, SWITCHBOARD: telephone speech corpus for research and development, in: ICASSP, Vol. 1, 1992, pp. 517–520.
- [20] A. Stupakov, E. Hanusa, J. Bilmes, D. Fox, COSINE - A Corpus of Multi-Party COnversational Speech In Noisy Environments, in: ICASSP, 2009.
- [21] M. Seltzer, Microphone Array Processing for Robust Speech Recognition, Ph.D. thesis, Carnegie Mellon University (2003).
- [22] T. Sullivan, Multi-Microphone Correlation-Based Processing for Robust Automatic Speech Recognition, Ph.D. thesis, Carnegie Mellon University (1996).
- [23] A. Subramanya, J. Bilmes, Virtual evidence for training speech recognizers using partially labeled data, in: HLT, 2007.
- [24] A. Subramanya, C. Bartels, J. Bilmes, P. Nguyen, Uncertainty in training large vocabulary speech recognizers, in: ASRU, 2007.
- [25] A. Subramanya, J. Bilmes, Applications of virtual-evidence based speech recognizer training, in: Interspeech, 2008.
- [26] P. Boersma, Praat, a system for doing phonetics by computer, *Glott International* 5 (9/10) (2001) 341–345.
- [27] C. M. University, cmudict0.7a, <https://cmusphinx.svn.sourceforge.net/svnroot/cmusphinx/trunk/cmudict/cmudict0.7a> (2008).
- [28] E. Shriberg, Spontaneous speech: How people really talk and why engineers should care, in: EUROSPEECH, 2005.
- [29] FLAC - Free Lossless Audio Codec, v1.1, <http://flac.sourceforge.net/>.
- [30] D. Heckerman, C. Meek, Models and selection criteria for regression and classification, in: UAI, 1997.
- [31] M. Zimmermann, A. Stolcke, E. Shriberg, Joint segmentation and classification of dialog acts in multiparty meetings, in: ICASSP, 2006.