COSINE - A CORPUS OF MULTI-PARTY CONVERSATIONAL SPEECH IN NOISY ENVIRONMENTS

Alex Stupakov, Evan Hanusa, Jeff Bilmes, and Dieter Fox

Department of Electrical Engineering, University of Washington, Seattle, WA, USA
{stupakov,hanusaem,bilmes}@ee.washington.edu, fox@cs.washington.edu

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. We explain the motivation for creating such a corpus and describe the resulting audio recordings and transcriptions that comprise the corpus. These recordings include a 4-channel array and close-talking, far-field, and throat microphones on separate synchronized channels, allowing for unique algorithm research.

Index Terms— Microphone arrays, speech recognition, multi-party, noisy speech corpora, throat microphone

1. INTRODUCTION

In many applications, speech recognition systems must be robust to the presence of background noise in the environment. These applications are numerous, and include dictation software or speech-based human-computer interfaces to be used in noisy environments; speech recognition of cellphone or air traffic control conversations; speech recognition or keyword search of noisy audio such as recordings of interviews; and voice commands used by soldiers, firefighters, law enforcement officials or disabled individuals to interact with assistive devices in the presence of background noise.

When training a speech recognition system that must work in noisy environments, two types of effects must be overcome: the presence of additive or convolutional noise as well as reverberation, and the effect of the noisy environment on the nature of the speech (Lombard effect). The methods used to mitigate these effects fall into three main categories [1]. 1) Noise cancellation or reduction can be performed on the audio signal before passing it into the speech recognizer. 2) Noise-robust feature extraction methods can be used to gain performance over standard MFCC features [2]. Mean/variance normalization and feature smoothing [3] and a variety of other feature cleaning/enhancement techniques show improvement over standard MFCC/PLP features [4, 5, 6]. 3) The acoustic models can be compensated for the noisy environment either by training on a combination of clean and noisy speech, or by using speech recorded in the desired noisy environment to adapt an existing acoustic model. The use of training audio that exhibits the Lombard effect has also been shown to improve the performance of speech recognition systems [7].

This work was supported in part by DARPA's ASSIST Program (contract number NBCH-C-05-0137) and an ONR MURI grant (No. N000140510388).

The performance of practical speech recognition systems relies on the availability of training data that was recorded in similar conditions to the audio being recognized. A broad selection of speech corpora exists; however, few of them provide ideal training data for performing automatic speech recognition (ASR) on in-situ conversational speech recorded in noisy environments.

Many corpora have been developed for studying algorithmic improvements that provide ASR performance increases. For example, the AURORA database [8] contains noisy and clean recordings of spoken digits. Other corpora have been designed to capture the Lombard effect, including UT-Scope [9] and the Albayzin Spanish-language corpus [10]. The ICSI [11] and AMI [12] 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 [13] and the CIAIR Japanese corpus [14], which also includes dialog recordings. There are also databases which capture the effects of specific types of distortion, for example, telephone channels in Switchboard [15].

Our goal was to create a corpus that brings together many of the elements that make each of these corpora useful: the presence of various levels and types of background noise, recordings of Lombard speech with and without the background noise, multi-microphone recordings of the same speech (including a microphone array), and spontaneous multi-person in-situ conversations.

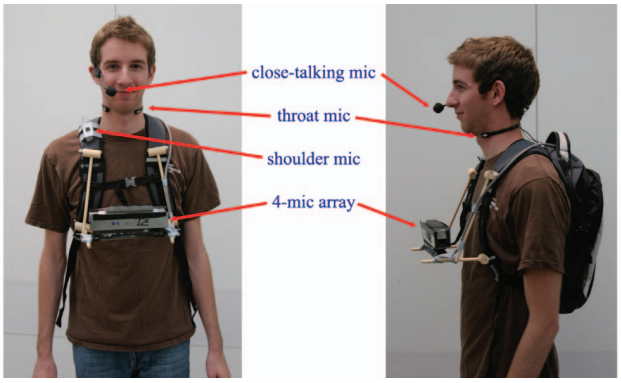
This paper describes such a corpus that has been recently collected. Many considerations motivated the design of the hardware and data collection practices. The corpus contains multi-party conversations about everyday topics in a variety of noisy environments. These noisy environments range in both noise type and intensity. Additionally, the speech is recorded in the environment in which the noise occurs, rather than having the noise added later. The portability of the recording devices allows for the speech to be recorded in-situ, rather than making the recordings in a studio, which affects the speakers' comfort and speech patterns.

2. RECORDING EQUIPMENT

A portable recording setup with seven microphones was designed for this data collection. It consists of a lightweight backpack, an array of four electret microphones (spaced 3 cm apart) 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.

A broad range of audio quality is captured by the various microphones in the system, whose audio streams are all recorded onto separate but synchronized channels. All of the array channels are

recorded on one device, achieving sample-level synchronization. The other three channels are recorded on a second device, synchronized with the first to within 10ms. Within a conversation, the recordings of all participants are synchronized to within 100ms. The synchronization allows this data to be used in research on dialog acts and conversational dynamics, such as described in [16]. The close-talking microphone records high quality audio of the individual's speech, the throat mic records distorted speech with near-ideal background noise rejection, and the shoulder and array microphones record significant background noise, including speech of other individuals.



(a) Front view (b) Side view

Fig. 1. The wearable recording system

3. RECORDING SESSIONS

Paid volunteers participated in multi-person recording sessions that lasted between 45 minutes and 1.5 hours. The breakdown of number of people per session is: 2 people: 13%, 3 people: 19%, 4: 42%, 5: 3.5%, 6: 19%, 7: 3.5%. 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. Six sessions were transcribed fully, and four were transcribed partially (not all speakers were transcribed). The total length of transcribed audio is 36.5 hours, of which 26.7 hours is speech and 9.8 hours is non-speech. The total length of audio in the 23 non-transcribed sessions is 145 hours.

3.1. Speakers

All the speakers whose voices were recorded are fluent (but not necessarily native) English speakers. There were 91 unique speakers. Of them, 59 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 will be released along with the audio recordings and transcripts. The number of sessions recorded by speakers of each gender is shown in Table 1.

	Male	Female	Total
Transcribed	15	22	37
Untranscribed	38	58	96
Total	53	80	133

Table 1. Gender of recorded speakers

3.2. Noise types

The recordings were made indoors and outdoors, on and near the campus of the University of Washington in Seattle, WA, 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. 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.

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: a) fully-labeled (FL) - transcribers mark the precise beginnings and ends of words, b) sequence-labeled (SL) - transcribers mark the beginning and end of a phrase and then transcribe only the sequence of words, and c) a technique introduced in [17] called partially-labeled (PL) - the word sequence is transcribed, and an identifying mark is placed somewhere within each word.

As shown in [18], 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 performance over both (53.1 WER on SWB_eval01 for FL, 53.9 for SL, and 51.8 WER for PL). While recent results have shown that the PL approach can be approximated by using a two-pass training strategy [19], PL can benefit from human annotators since they are immune to OOVs and disfluencies that become more frequent as the colloquial nature of the speech in the corpus, such as our own, 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 2 shows an example of a PL annotation with a privacy deletion (discussed in section 5.3).

5. ORTHOGRAPHIC TRANSCRIPTION

The transcription of the corpus was done by a group of 3rd and 4th year Linguistics undergraduate students using the Praat program [20]. All transcribers are native English speakers.

To ensure consistency of transcription, an annotation guide was created. A wiki was also established, allowing the transcribers to standardize their transcriptions and to learn from each other. The wiki contains 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

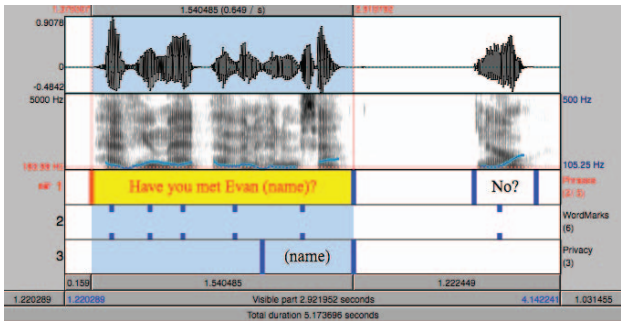


Fig. 2. Example annotation in the Praat program. Tiers 1, 2, and 3 show phrases, PL word marks, and privacy regions, respectively.

concatenation of the pronunciations of their constituents. The resulting list has over 400 words, and will be included in the corpus release. The wiki was frequently updated, and served as a 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 transcriber 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%.

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 [12].

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. Before the submission of a completed annotation segment, a spellcheck ensured that all words are in the cmudict0.7a dictionary [21] or one of the word lists in the wiki. Lastly, 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.

5.2. Transcription Guide

The words in each region are transcribed in American English. Transcribers had access to the cmudict0.7a dictionary, and were able to search through it to determine the presence or absence of a word. Punctuation (period, comma, question mark, exclamation mark) is included in the transcriptions.

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. Any word that is completely unintelligible is marked with a "+". A sequence of unintelligible words is marked with repeated "+" symbols, and each "+" is marked with a single word interior mark.

Once the phrase has been transcribed, words that are "mispronounced" are marked with an asterisk ("*"). Whether or not a word is "mispronounced" is subjective, and was 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, the transcriber is instructed to do the following: if the speaker always mispronounces the word in the same way, this pronunciation is added to the wiki as an alternate pronunciation of this word, and the word is not marked with a "*". 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 derived from the data. Words marked with a "*" can either be ignored for this purpose or can be included as the researcher desires. Mispronounced words are not excluded completely because the transcripts will still be useful for training of language models or other text-based research, which is also very important in speech recognition [22].

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 tilde (~), and their pronunciations are added to the dictionary appendix. Some examples are "NAFTA~" and "NASA~".

Additional special symbols are introduced: "#" is used to indicate whistling, coughing, sneezing, non-verbal singing, or miscellaneous vocal noise. "$" is used to indicate laughing. It can be a standalone symbol or can be appended to a word to indicate laughing and talking simultaneously. "@" is used to indicate a foreign word. It can replace an unknown foreign word or be appended to a word (for example: "bonjour@"). "-" is used to indicate a disfluency or discontinuity at the beginning or end of a word. Singing or melodic speaking is not annotated in any special way. Commonly occurring suffixes ('ve, 'll, 's) are transcribed as separate words if they are not part of a common compound (such as "I've"), because full words containing these suffixes (such as "Jenny'll") are unlikely to be found in a dictionary. If necessary, these suffixes can be joined with the preceding word using simple post-processing.

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.

6. RELEASE

The final release of the corpus will contain all of the original recorded audio (excepting any privacy-related deletions) stored in the FLAC compressed lossless audio format [23], the transcriptions, and all non-privacy-sensitive subject information. The corpus will be made available online[1] to speech researchers free of charge.

7. CONCLUSION

The COSINE corpus will 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 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.

8. REFERENCES

[1] Y. Gong, "Speech recognition in noisy environments: a survey," *Speech Communication*, vol. 16, no. 3, pp. 261–291, 1995.

[2] W. Guo, L. Zhang, and B. Xia, "An auditory neural feature extraction method for robust speech recognition," in *ICASSP*, 2007.

[3] C.P. Chen, *Noise Robustness in Automatic Speech Recognition*, Ph.D. thesis, University of Washington, 2004.

[4] M.B. Islam, H. Matsumoto, and K. Yamamoto, "An improved mel-Wiener filter for mel-LPC based speech recognition," in *Interspeech–ICSLP*, 2006.

[5] W. Lim, C.W. Han, J.W. Shin, and N.S. Kim, "Cepstral domain feature compensation based on diagonal approximation," in *ICASSP*, 2008.

[6] D. Yu, L. Deng, J. Droppo, J. Wu, Y. Gong, and 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, and D. Paul, "Multi-style training for robust isolated-word speech recognition," in *ICASSP*, 1987, vol. 12, pp. 705–708.

[8] H.G. Hirsch and D. Pearce, "The AURORA experimental framework for the performance evaluation of speech recognition systems under noisy conditions," in *ASR2000-Automatic Speech Recognition: Challenges for the new Millenium ISCA Tutorial and Research Workshop (ITRW)*, 2000.

[9] V.S. Varadarajan, J.H.L. Hansen, and 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.

[10] A. Moreno, D. Poch, A. Bonafonte, E. Lleida, J. Llisterri, J.B. Marino, and C. Nadeu, "Albayzin speech database: Design of the phonetic corpus," in *EUROSPEECH*, 1993, pp. 175–178.

[11] A. Janin, D. Baron, J. Edwards, D. Ellis, D. Gelbart, N. Morgan, B. Peskin, T. Pfau, E. Shriberg, A. Stolcke, and C. Wooters, "The ICSI Meeting Corpus," in *ICASSP*, 2003, vol. 1, pp. 364–367.

[12] J. Carletta, S. Ashby, S. Bourban, M. Flynn, et al., "The AMI meeting corpus: a pre-announcement," *Lecture notes in computer science*, vol. 3869, pp. 28, 2006.

[13] B. Lee, M. Hasegawa-Johnson, C. Goudeseune, S. Kamdar, S. Borys, M. Liu, and T. Huang, "AVICAR: audio-visual speech corpus in a car environment," in *ICSLP*, 2004.

[14] N. Kawaguchi, S. Matsubara, H. Iwa, S. Kajita, K. Takeda, F. Itakura, and Y. Inagaki, "Construction of speech corpus in moving car environment," in *ICSLP*, 2000.

[15] J.J. Godfrey, E.C. Holliman, and J. McDaniel, "Switchboard: telephone speech corpus for research and development," in *ICASSP*, 1992, vol. 1, pp. 517–520.

[16] M. Zimmermann, A. Stolcke, and E. Shriberg, "Joint segmentation and classification of dialog acts in multiparty meetings," in *ICASSP*, 2006.

[17] A. Subramanya and J. Bilmes, "Virtual evidence for training speech recognizers using partially labeled data," in *HLT*, 2007.

[18] A. Subramanya, C. Bartels, J. Bilmes, and P. Nguyen, "Uncertainty in training large vocabulary speech recognizers," in *ASRU*, 2007.

[19] A Subramanya and J. Bilmes, "Applications of virtual-evidence based speech recognizer training," in *Interspeech*, 2008.

[20] Paul Boersma, "Praat, a system for doing phonetics by computer," *Glot International*, vol. 5, no. 9/10, pp. 341–345, 2001.

[21] Carnegie Mellon University, "cmudict0.7a," `https://cmusphinx.svn.sourceforge.net/svnroot/cmusphinx/trunk/cmudict/cmudict0.7a`, 2008.

[22] E. Shriberg, "Spontaneous speech: How people really talk and why engineers should care," in *EUROSPEECH*, 2005.

[23] "FLAC - Free Lossless Audio Codec, v1.1," `http://flac.sourceforge.net/`.

[1] Please web-search for "cosine speech corpus" to find the corpus's current residence.