VLSP 2011

New Tools and Methods for Very-Large-Scale Phonetics Research Workshop

January 29-31, 2011
The University of Pennsylvania

Workshop Program and Abstracts/ Papers

Mining a Year of Speech

John Coleman[1], *Mark Liberman*[2], *Greg Kochanski*[1], *Lou Burnard*[3], *and Jiahong Yuan*[2]

[1] Phonetics Laboratory, University of Oxford, England
[2] Phonetics Laboratory, University of Pennsylvania, USA
[3] formerly Computing Services, University of Oxford, England
{john.coleman, greg.kochanski}@phon.ox.ac.uk, {myl@cis, jiahong@ling}.upenn.edu,
lou.burnard@retired.ox.ac.uk

Abstract

The availability of large text corpora has revolutionized linguistics and is of great value in many other areas of scholarship. Our "Mining a Year of Speech" project, funded by the transatlantic "Digging into Data" competition, aims to do the same for spoken language. We present a new generation of speech corpora, characterised by aggregation of datasets, annotated using forced alignment and exposed for public use in standard formats across multiple sites.
Index Terms: speech corpora, aggregation, forced alignment

1. Introduction

Spoken language, particularly ordinary conversation, has a particular importance in the study of language (e.g. [1], [2]). It predominates in all cultures and times, and is acquired naturally rather than through formal education, but for practical reasons spoken language research has tended to use relatively small datasets, usually elicited in specific, controlled situations and recorded in a studio. However, with the advent of huge amounts of accessible digitized speech (including speech that accompanies digital video) and significant advances in automatic alignment techniques, it is now possible to combine the benefits of working on a very large scale with the fine detail and naturalness of speech recorded outside the studio.

For corpora, large size is important because many aspects of speech and language are characterized by huge numbers of possible alternatives, each of which is individually rare (see also [3]). Language is characterized by a hierarchy of improbability: to find examples of more complex structures, or combinations of structural, sociolinguistic or contextual variables larger corpora are required. For instance, most phonemes of a language occur within a minute of speech, and finding pronunciations of common words requires about an hour of speech. The 10.4-million word spoken part of the British National Corpus (henceforth "Spoken BNC") contains c. 64,000 distinct word-forms, 85% of which occur only once (e.g. abhorred), whereas just five words (the, I, it, you, and, I) each occur over 200,000 times. As a result, obtaining even a moderate number of samples of most words of interest requires starting with a very large collection.

To understand the phonetics of *combinations* of words, an appropriate sample of word pairs requires a corpus containing hundreds of hours of speech. For example, in another research project that we are just beginning, we will study the pronunciation of word-joins, mining the Spoken BNC for evidence of assimilation vs. non-assimilation of word-final nasals, as in e.g. "seem to" vs. "seen to". While "seem to" occurs 310 times in the Spoken BNC, there are only 12 instances of "seen to": if the corpus were smaller, there might be none!

The complete collection of corpora we are working on in the Mining a Year of Speech project will include about 9000 hours (100 million words, or 2 Terabytes) of speech in various American and British dialects of English, derived from the Linguistic Data Consortium, the British National Corpus, and other existing resources. While even this is not large enough for every purpose, it will permit the extraction of subsets appropriate for addressing many questions: phonetic, linguistic and otherwise. The project addresses the challenge of providing rich, intelligent data mining capabilities for a substantial collection of spoken audio data, applying state-of-the art techniques to offer sophisticated, rapid, and flexible access to a set of richly annotated corpora. This is at least ten times more data than has previously been used by researchers in fields such as phonetics, linguistics, or psychology, and more than 100 times common practice in spoken language research.

It is of course impractical for a researcher to listen to a year of audio - and one anticipates even larger corpora in the near future - in order to search for certain words or phrases, or to manually measure the resulting data. However, by using *forced alignment* to add rich annotation to large audio corpora, the task of *finding* relevant data could take just a few seconds. Though our experience and research interests happen to be focussed on phonetic matters such as intonation, pronunciation differences between dialects, and dialogue modeling, the text-to-speech alignment and search tools used by the project will open up this 'year of speech'

2. The datasets

For the 'year of speech' we shall not create a single new corpus at one site, but shall expose a collection of corpora (we use the collective term "grove" of corpora, suggested in personal conversation by Sebastian Rahtz) under common indexing standards to a cross-searching front-end. We are aggregating a collection of transcribed audio corpora which we already have at Penn, Oxford and the British Library, with metadata and other annotations. One portion will be speech corpora published by the Linguistic Data Consortium (LDC), whose catalog now includes about 4,200 hours (almost six months) of recorded and transcribed American English speech, and more than 2,000 hours of as-yet unpublished material, including broadcast conversations such as talk shows, Supreme Court oral arguments, political speeches and debates, and audio books. For British English, the largest collection of transcribed spoken audio in existence is the spoken part of the British National Corpus, of which we have 7.4 million words recorded as audio (i.e. only three quarters of the 10 m words of transcribed speech).

Although the transcription details vary from corpus to corpus, they are typically orthographic transcriptions at the word level, with linguistic annotations and metadata represented as XML data structures. Segment-level labelling is derived as a by-product of forced alignment of orthography to audio, as illustrated in Figure 1. To align the transcriptions with the audio, we employ an automatic speech-to-phoneme alignment system, the Penn Phonetics

Lab Forced Aligner [4], based on the HTK speech recognition toolkit ([5]). (To transcribe and align 1,200 hours of speech manually could easily take at least 240,000 hours of researcher time, c. 120 person-years.)

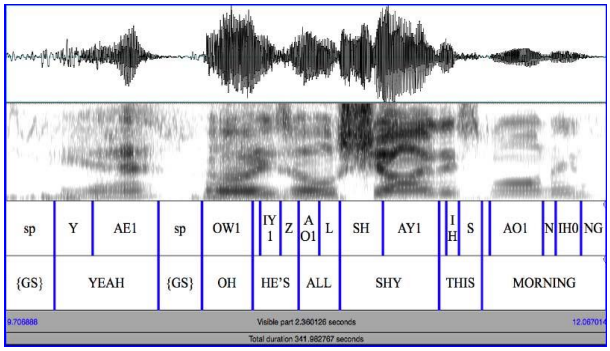

Figure 1: *Example of phonemic and orthographic transcriptions aligned with a BNC speech sample.*

As with the existing text version of the British National Corpus, others may wish to create and provide other tools to search or perform other processes on the exposed data.

2.1. American English sound files at the LDC

The transcribed American English speech in the LDC's current catalog includes about 2,240 hours of two-party telephone conversations, 260 hours of task-oriented dialogs, 100 hours of group meetings, 1,255 hours of broadcast news, 30 hours of voice mail, and 300 hours of read or prompted speech. As-yet unpublished LDC data includes about 1,000 hours of broadcast conversations (such as talk shows), and about 5,000 hours of U.S. Supreme Court oral arguments. The process of aligning all of this at the word and phonetic-segment level is under way, using the automatic forced-alignment techniques developed at the Penn Phonetics Lab. We are in the process of adding a large sample of political speeches and debates, and a collection of open-access audio books from librivox.org.

2.2. British English sound files in the BNC

The Spoken BNC uniquely combines informality with size. It is the largest and most comprehensive snapshot of "spoken language in the wild" ever collected: over 1200 hours. (The design of the corpus and the methodology of collection and mark-up have been described in various earlier publications, especially [6], [7], [9].) It was collected in c. 1991-2 for Longman, who contributed it to the BNC Consortium. For the "demographic" half of the corpus, volunteers recruited by a market research company (BMRB) carried a Sony Walkman tape recorder around with them for several days, yielding a cross-section of everyday speech, including lunchtime conversations, discussions of boyfriends, and dog-directed speech. About half of the Spoken BNC comprises such conversations which are unstructured, wide-ranging, often involving multiple people in very different kinds of speech situation. The other half is more formal, but mostly unscripted speech, such as interviews, informal meetings, and religious services recorded in a wide variety of predefined social contexts. The recorded audio was transcribed by professional audio typists at the word level, to which rich linguistic annotation and metadata was subsequently added.

Because of its size and nature, the Spoken BNC will be an exceptionally valuable resource for studies of English phonetics and phonology. Sociolinguists and others have collected unscripted corpora (typically sociolinguistic interviews) but all, we believe, are substantially smaller than the Spoken BNC. The largest comparable transcribed and aligned dataset is the Switchboard corpus [10] which amounts to just 300 hours of audio -- about a quarter the size of the Spoken BNC. Other comparable, noteworthy corpora are the Buckeye Corpus of American English [11] and the ONZE corpus [12], which has 200 hours of time-aligned phrase-level transcriptions. Among the largest British sociolinguistic corpora, the York Corpus [13] (Tagliamonte 1998) contains c. 100 hours of transcribed sociolinguistic interviews, and the Newcastle Electronic Corpus of Tyneside English [14] contains c. 10 hours of transcribed audio.

As part of Oxford University and the British Library's institutional contributions to the project, the British Library Sound Archive (where the tapes are deposited) has now completed the digitization of the 1,213 90-minute tape recordings from the Spoken BNC. Based on the text of the BNC, we have augmented the dictionary used by the Penn Phonetics Laboratory Forced Aligner with phonemic transcriptions appropriate to a variety of British English pronunciations, including all the oddities such as rare names, *hapax legomena*, truncated/incomplete words, etc. and are now aligning every word and phoneme with the corresponding portion of the audio recording. The main dimensions of variation in the transcriptions are: (1) British English [ɒ] vs. [ɑ] (a contrast that is neutralized to [ɑ] in American English); (2) Northern British English [ʊ] vs. Southern British English [ʌ]; (3) postvocalic [r] or its absence (varies across British English); some systematic stress differences.

Forced alignment will allow easy access to desired words and phones. Only a small part of the audio transcribed for the BNC has previously been generally available to researchers (the COLT corpus [16], a sub-corpus of the Spoken BNC); the large remainder of the Spoken BNC is currently not easily available to most researchers as the tapes can only be audited by visiting the BL Sound Archive in person. Furthermore, searching the tapes for samples of interest is prohibitively time-consuming. This project aims to provide much better access to the anonymized audio recordings.

With the assistance of an academic visitor, Céline Poudat, we examined the deposited paperwork relating to the recordings, discovered how the recording numbers relate to the XML database, and resolved the IP/licensing conditions for use and eventual release of this data, clarifying the legal and technical feasibility of joining together the digitized audio recordings with the transcribed corpus. This confirmed the legal and technical feasibility of releasing the digitized recordings with the transcriptions.

The original publication terms set out in the spoken permissions request letter require us to blank out names and other speaker-specific information from the recordings, because the speakers were promised anonymity. The necessary deletions are already marked in the text part of the Spoken BNC using XML <gap> tags; we will simply mute the audio corresponding to the <gap> tags. Manual checking will ensure in due course that such deletions have been correctly applied. A close variant of this procedure was previously employed in the fraction of the BNC already published as the COLT corpus. The literature on voice identification (e.g. [17], [18]) indicates that release of the audio with appropriate deletions will not allow identification of the Spoken BNC volunteers, especially since comparison recordings should not be easily available after 17 years.

The final result will be a multi-million-word transcribed audio corpus where researchers can locate the speech corresponding to points in the text. The Spoken BNC will then be a world-class resource for sociophonetics, phonetics

and phonology research, or any area of study where ordinary spoken language is of interest (e.g. ethnomethodology).

2.3. Alignment of transcriptions to audio

In 2008 we conducted a 5-month pilot project to assess the application of the Penn Phonetics Lab Forced Aligner to a sample (3%) of the BNC recordings. This pilot determined that the audio quality is sufficient for use with the Penn Phonetics Lab Forced Aligner and quantified the computational requirements. Although the Penn Aligner's acoustic models have been trained on American English speech, its application to British English dialects was surprisingly successful: in our sample, 83% of the computed phoneme boundaries were located within 2 seconds of their correct position. Although this is not very accurate, it is good enough for a researcher to be able to select almost any word in the Spoken BNC, look up its position, and automatically display the relevant portion of audio (probably the correct utterance) on the screen. This capability will make over 6 million spoken words available for analysis. An additional goal is to determine a running confidence measure for the accuracy of alignment across the whole corpus ([19] reports on work in progress). For a smaller but still substantial portion, we can automatically locate individual phonetic segments within the Spoken BNC, since 24% of the segment boundaries were within 20 ms of expert human labels. This is not a large fraction, but when one boundary is accurate, almost all the boundaries within 2 seconds are also accurate. This fact should allow us to tell whether each given region is aligned well.

Compared to the usual quality measures of speech recognition systems, statistics such as the above are distinctly unimpressive. But it is important to note that (a) alignment accuracy measures are usually intended to be reports of *overall best case performance*, whereas our statistics were compiled to estimate a *lower* bound on the system's expected performance. "83% accurate within 2 s" means that one can roughly locate most of the instances of words or phoneme patterns one is searching for, within about a sentence or so. If one accepts the need to home in on the exact portion manually, this is still good enough to navigate the corpus in order to mine it for examples manually. "24% accurate within 20 ms" means that on average one can reliably find ¼ of the desired material completely automatically, if one has a running confidence measure. Roughly speaking, this means that about ¼ of the Spoken BNC (c. 1.9 million words) is accurately labelled by the automatic forced-alignment software.

Despite this qualified optimism, the Spoken BNC is a challenging corpus for speech alignment. It has background noises ranging from televisions to birdsong and traffic, it covers a wide range of British English dialects, and the environmental acoustics varies from reverberant rooms to outdoors. Although the transcriptions are mostly accurate, they do not capture all filled pauses, repairs, and other conversational phenomena, and do not precisely represent simultaneous speech by several people. Although HMM-based forced alignment works well on read speech and short sentences, the alignment of long and spontaneous utterances remains a challenge. Spontaneous speech contains filled pauses, disfluencies, errors, repairs, and deletions that are often omitted in the transcripts, and pronunciations in spontaneous speech are much more variable than read speech. In our 2008 pilot project, we found that erroneous alignments could be reduced by adapting the silence and noise models of the Penn Phonetics Lab Forced Aligner to the BNC audio data. We are exploring the importance of modelling the background noise in between speech in improving the alignment of long and casual speech, and of adapting models to different speakers. Another type of error we have seen is that some words are extremely long in the alignment results. This usually occurs when there is long speech-like background noise surrounding the words. This type of error can be reduced by introducing constraints on word or phone duration.

3. Outputs

In addition to the many audio files comprising the 'year of speech', we shall create a new, Extended release of the BNC, which provides an additional structure of timing information to the existing text transcriptions. The transcriptions and timing information will be released in XML form. We intend to make it available under the same license and via the same arrangements as the XML text version of the BNC. The extended XML data will contain beginning and ending times of words and phonemes, along with audio recording identifiers.

This dataset is big enough to demonstrate what can be achieved using large amounts of data. It is several hundred times larger than the largest datasets previously used in research of this type; and just as important, it is much more diverse. Both the size and the diversity will make new kinds of research possible. But the most important thing of all is that the framework and tools that we are developing can easily be applied to any additional material for which both recordings and standard orthographic transcriptions exist. As a result, others will be able to apply our methods to new datasets of interest to them, including (for example) sociolinguistic interviews, oral histories, courtroom recordings, political speeches and debates, doctor-patient interactions, and so on. As licensing and other constraints allow, we hope to support scholars from other fields by extending the range of corpora included in the collection as far as possible. Since it will be a 'grove' of corpora, distributed across various centres but exposed according to common standards, we hope that others may wish to add to the collection so that we may all benefit from a resource greater than any of us might individually create.

4. Acknowledgements

We wish to acknowledge the support of JISC (in the UK) and NSF (in the USA) for their joint support of Mining a Year of Speech, under the internationally-coordinated Digging into Data programme. Digitization of the Spoken BNC was funded by the John Fell OUP Research Fund, the British Library and the University of Oxford Phonetics Laboratory; and our pilot project was also funded by the John Fell Fund. We thank all of these funders.

5. References

[1] de Saussure, F. Cours de linguistique générale, ch. 2, Payot, 1916.
[2] Abercrombie, D. "Conversation and spoken prose", in Studies in Phonetics and Linguistics, 1-9, Oxford UP, 1965.
[3] Kochanski, G., Shih, C., and Shosted, R. "Should corpora be big, rich or dense?", submitted to this conference.
[4] Yuan, J. and Liberman, M. "Speaker identification on the SCOTUS corpus". Proceedings of Acoustics '08, 2008.
[5] Young, S., Evermann, G., Gales, M. and 9 others, The HTK Book (for HTK Version 3.4). Online: http://www.ee.uwa.edu.au/~roberto/research/speech/local/htk/htkbook.pdf, accessed on 30 Nov 2010.
[6] Crowdy, S. "Spoken Corpus Design", Literary and Linguistic Computing, 8(4):259-265, 1993.
[7] Crowdy, S. "The BNC spoken corpus", in [8], pp. 224-234.
[8] Leech, G., Myers G. and Thomas, J. [Eds], Spoken English on Computer, Longman, 1995
[9] Aston, S. and Burnard, L. The BNC Handbook, Edinburgh UP, 1998.
[10] Godfrey, J. J., Holliman, E. C. and McDaniel, J. "SWITCHBOARD: Telephone speech corpus for research and development." Proc. ICASSP-92, 517–520, 1992.
[11] Pitt, M. A., Johnson, K., Hume, E., Keisling, S., and Raymond, W. "The Buckeye corpus of conversational speech", Speech Communication 45:89-95, 2005.
[12] Gordon, E., Maclagan, M. and Hay, J. "The ONZE Corpus", in [15], 82-104.
[13] Tagliamonte, S. "*Was/were* varation across the generations: View from the city of York." Language Variation and Change 10(2):153-191, 1998.
[14] Allen, W. H, Beal, J. C., Corrigan, K. P., Maguire, W., and Moisl, H. L. "A linguistic time capsule: the Newcastle Electronic Corpus of Tyneside English", in [15], 16-48.
[15] Beal, J. C., Corrigan, K. P. and Moisl, H. [Eds], Models and Methods in the Handling of Unconventional Digital Corpora, vol. 2, Palgrave, 2007.
[16] Haslerud, V. and Stenström, A.-B. "The Bergen Corpus of London Teenager Language (COLT)", in [8], pp. 235-242.
[17] Campbell, W.M., Campbell, J.P., Reynolds, D.A., Singer, E., and Torres-Carrasquillo, P.A. "Support vector machines for speaker and language recognition." Computer Sp. Lang. 20(2-3): 210-229, 2006.
[18] Kenny, P., Ouellet, P., Dehak, N., Gupta, V., and Dumouchel, P. "A Study of Interspeaker Variability in Speaker Verification." IEEE Trans. on ASLP, 16 (5):980-988, 2008.
[19] Baghai-Ravary, L., Grau, S. and Kochanski, G. "Detecting gross alignment errors in the Spoken British National Corpus", submitted to this conference.