NSF Project IIS-0535168: Separating Speech from Speech Noise Annual Report 2007

Daniel P.W. Ellis · Columbia University
Pierre Divenyi · East Bay Institute for Research and Education
Barbara Shinn-Cunningham · Boston University
Deliang Wang · Ohio State University

Jan 16, 2008

1 Activities & Findings

1.1 Research Activities

This year we continued our multisite collaboration aimed at improved understanding and modeling of how listeners handle noisy or corrupt speech.

On the **perception** side, we have continued experiments on how listeners perceive distorted and incomplete speech tokens, including the use of our spondee/foil paradigm, in which listeners are played tokens derived from two-syllable compound words which are either relatively familiar spondees (e.g. "bankroll"), or pairs of monosyllable words less likely to be encountered together, but with the same initial consonant-vowel, and final vowel-consonant (e.g. "band hole"). Our manipulations consist of replacing parts of the original speech with short acoustic excerpts that provide reduced information, for instance by replacing a 200 ms chunk with 200 ms of white noise. In the spondee/foil paradigm, we generally replace a section of the speech centered on the consonant cluster between the two syllables, which tends to mask the distinction between spondee and foil. Increasing the temporal width of the replacement, and/or reducing the amount of information from the original speech preserved in the replacement, will lead to

a greater probability of misperceived words, or confusions between spondee and foil.

In addition to varying the duration of the replacements, we have used various kinds of replacements that preserve different aspects of the original speech. Silent gaps or unmodulated noise fillers provide almost no information about the original speech (although noise allows the perceptual system to hypothesize that speech was in fact present but was obliterated by the louder imposition, leading to better intelligibility than silent gaps – the well-known "phonemic restoration" effect). More information can be provided by amplitude-modulating the noise to follow the temporal energy contour of the speech that is being replaced, and/or using a pulse-sequence (buzz) whose period reflects the pitch track of the original speech. The perceived spatial location of the replacements can also be manipulated by introducing a small delay between the signals presented to each ear such that the replacements may occur at the same location as the original speech, or spatially separate from it.

Finally, where the replacements are modulated to carry amplitude and/or pitch information from the original speech, this information can be either correct or inconsistent (e.g. time-reversed, or taken from a completely different speech prototype). Results of this range of experiments are summarized in the finding section below.

On the **modeling** side, we have pursued a number of computational techniques aimed at reproducing the abilities of listeners to 'hear out' speech in noisy and distorted circumstances. In terms of directly modeling the spondee/foil experiments, we have looked at articulatory models, such as the "Task Dynamics" model of Saltzman and Munhall [6], as a representational space in which misperceived word sequences may be most closely reconciled to the original speech. We use a task-dynamics speech synthesizer [3] to create approximate articulator sequences corresponding to the original and perceived utterances (translated into phonemes via a dictionary), then attempt to account for the perceptual impacts of modifications as "pushing" the articulator motions that may be being perceived by the listener towards the gestures corresponding to the words reported as heard.

We have also continued developing our models of identifying and separating signals across a broad range of distorted and masked speech conditions. We have paid particular attention to the problem of segregating unvoiced speech sounds, which by their nature cannot be isolated based on the pitch cues that have been widely explored for the separation of vowels from interference, and therefore present a much greater challenge both to computational models, and, by implication, for listeners. Note that the noise inserts in the modified speech tokens lack

pitch and thus share many of the characteristics of unvoiced speech; however, in the absence of efforts to integrate the noise with the speech such as consistent amplitude modulation, listeners do not attempt to interpret them as unvoiced speech sounds, but realize that they are external, independent signals that interfere with the speech.

Since they exhibit minimal signal structure, unvoiced sounds may be identified only on the basis of changes in spectral energy; our approach is to segment the initial acoustic mixture into a collection of locally-contiguous energy regions, each of which may be a part of the target speech or which may result from interference. The correct assignment of these regions to target or interference must depend on information beyond the local signal structure; we use a Bayesian classifier to decides whether segments are dominated by unvoiced speech or interference on the basis of acoustic-phonetic features derived from these segments. The proposed algorithm, together with our previous approach to voiced speech segregation, leads to a system that segregates both unvoiced and voiced speech from nonspeech interference.

The use of prior knowledge of the structure of acoustic sources can be further exploited for acoustic scene analysis through a process of parameter fitting through which a constrained model is made consistent with an observed target-interference mixture. We have investigated this idea in two forms: For spatial information, we have looked at inferring between-channel (e.g. interaural) parameters for individual spatialized sources in the presence of reverberant interference. By assuming only weak constraints such as consistency of between-channel delay across frequencies, and consistency of between-channel level differences along time, we can develop a process of iterative re-estimation to identify the location and characteristics of multiple source – e.g. three sources in a two-microphone scenario – despite reverberant interference.

In the absence of usable spatial cues, it is still possible to separate the information of overlapping sources by using more detailed models of the source characteristics. For instance, if the listener has a strong idea of the kinds of sounds that are likely to emerge from a given speaker – a model of that speaker's voice – then these expectations can be used to differentiate between aspects of a sound mixture that have originated with that speaker, and other parts that are interference. We attempt to model this with speaker-specific speech models, representing the particular set of speech sounds a given speaker can emit. This approach has already been shown to be very effective at recognizing superimposed speech, even to the extent of exceeding human performance in some circumstances [5]. However, to avoid the restriction that the characteristics of the target speaker be known

in advance, we have investigated using a parametric model of individual speaker variation, trained from a collection of individual speech examples. During separation, the parameters that best match the target speaker are estimated, then the separation and parameter estimation is iterated to progressively refine both. We propose this as a rough computational analog of the way listeners can adaptively "tune in" to a particular voice in noisy situations.

As a further component of accurate and effective models of target speech signals, we have continued investigating the nature of pitch in speech. To determine the perceptually-salient level of detail, we conducted experiments in which the original pitch of real speech tokens was replaced by synthetic, simplified pitch contours based on different algorithms including a simple straight line per syllable. These resyntheses were compared with the original contours in a listening test to judge the naturalness of high-quality resynthesized speech.

1.2 Educational activities

The educational activities arising from this project include:

- Supervision of the five research assistants/graduate students and one undergraduate student directly involved in this project across the four partner institutions.
- The research has continued to provide input and material for various advanced courses taught by the PIs, including topics for projects. The work on pitch modeling, performed by an undergraduate, that has now led to a first-tier conference publication [4].
- One junior researcher spent two months at our collaborative lab in Paris supported by an OISE supplement to the project. This trip (detailed in a separate report) was of great educational benefit.
- Various substantial interactions between the researchers working at the different sites, including discussion between modeling engineers and experimental psychologists on how to account for results, and cross-participation on Ph.D. defense committees.

1.3 Findings

The findings from the research projects listed above include:

- When presenting incomplete spondee or foil word-pairs with variable amounts of deletion, correct recognition decreases as the amount of information removed increases, leading to worse discrimination between the two, and an increasing likelihood that subjects will choose the spondee (which has a much greater a-priori likelihood in general language) than the foil.

- When deleted regions are replaced by interference that carries some of the information of the original speech, such as its amplitude envelope or pitch track, this preserved information does improve intelligibility, but only if it is perceptually integrated with the undeleted speech. When perceptual integration is discouraged by presenting original speech and interference with different interaural timings (leading to the perception that they originate in different directions), the benefit of the preserved information is lost. In fact, preliminary results indicate that providing the correct amplitude envelope, but with interaural timing to locate it somewhere distinct from the original speech, can actually lead to worse intelligibility than using unrelated amplitude modulation. If this result is confirmed, it may be due to a process of across-object inhibition, where the perception of a pattern in one object actually makes it harder to perceive the same pattern in a second perceptually-distinct object [8].

- Modeling the perception of corrupted speech tokens based on fitting articulator-based gestures to the available information shows promise as a way to directly predict the kinds of confusions or misperceptions that listeners will make.

- Separation of unvoiced speech by machine has been shown through systematic evaluation to extract a majority of unvoiced speech without including a lot of interference. Our approach performs substantially better than existing techniques such as spectral subtraction [12].

- Rigorous statistical inference of the interaural parameters making up a source spatialization model is successful at locating, and separating, sources based only on the outputs two microphones, even when there are more than two sources present, and even in the presence of moderate reverberation. This

approach outperforms previous methods which are usually disrupted by reverberant reflections [2].

- Parametric speaker models can be used as models specific to the sound of a particular speaker's voice for the purposes of separating single-channel mixtures of two voices, without the need for prior knowledge of either speaker's voice. The parameters describing each speaker can be successfully inferred from the mixture, and the entire separation and parameter estimation process iterated to refine performance [11].

- Simple pitch contours consisting of per-syllable piecewise-linear segments are in many cases considered equally natural-sounding as the original pitch contour, in a listening test based on high-quality resyntheses. A previous pitch contour model achieved much lower scores [4].

2 Training and development

This year, the project has provided support for four current Ph.D. students and one research assistant who plans to enter a Ph.D. program in this area. In each case, the support has facilitated their continuing development into full participants in the research community, including studies, experiments, publications, and presentations. The project also enabled the supervision of one graduate student who has now gone on to enter a Ph.D. program at UC Berkeley.

3 Outreach

Researchers working on this project presented the ongoing results at a large number of international meetings and seminars including: Acoustical Society of America meetings (Salt Lake City and New Orleans), Association for Research in Otolaryngology Midwinter Meeting (Denver), Hearing Aid Developers' Forum (Oldenburg, Germany), International Conference on Cognitive and Neural Systems (Boston), UC Berkeley Hearing Seminar (Berkeley), Neuromorphic Engineering Workshop (Telluride), Workshop on Statistical Models for Speech and Audio (Radcliffe Institute, Cambridge), Boston University Hearing Reseach Center (Boston), Danish ASIP.NET Seminar (Copenhagen, Denmark), International Conference on Phonetic Science (Saarbrücken, Germany), IEEE International Conference on Acoustics, Speech and Signal Processing (Hawai'i).

From October 2006 to June 2007, DeLiang Wang visited Oticon A/S, a major hearing aid manufacturer located in Copenhagen, Denmark. He has collaborated with the Oticon signal processing group as well as the Oticon Eriksholm Research Center in an effort to evaluate the potential benefits of speech separation algorithms in improving speech intelligibility of hearing impaired listeners.

4 Contributions

4.1 Contributions to the principal disciplines

Results in speech perception have revealed and refined new and important phenomena in listeners' treatment of incomplete or corrupt speech tokens, including the complex interaction between phonemic restoration and perceptual organization into sources. Interpreting perceptual confusions in terms of inferred articulation offers a powerful framework for predicting the precise nature of misperceptions.

Computational modeling work has contributed important practical approaches to the separation of unvoiced energy, the localization and separation of overlapping voices in reverberation, the separation of previously-unknown voices in single-channel mixtures, and perceptually-adequate representation and synthesis of voice pitch contour.

4.2 Contributions to sister disciplines

Insofar as the two disciplines covered by this project, speech perception, and speech processing, are sisters, they have both contributed to the other. The speech perception experiments have provided specific data that can be used to constrain and refine computational models of speech perception in noise. The speech separation and recognition work has provided information such as fine time alignments needed to prepare the specific acoustic stimuli used in the perceptual experiments. Both threads are leading towards automatic speech separation systems that actually improve listeners' perception of speech with enormous implications for hearing instruments; Our ongoing relationships with hearing aid manufacturers will help realize these implications.

4.3 Contributions to human resources

The project has supported and provided for the supervision of four Ph.D. students, one pre-Ph.D. research assistant, and one undergraduate who has now entered a Ph.D. program. PIs have also contributed to the supervision and development of several other graduate and undergraduate students.

4.4 Contributions to educational infrastructure

The project has supported the continuing development of classroom and practical materials, many of which are made freely available over the web.

4.5 Contributions to wider aspects of public welfare

The motivation behind the speech separation project is to develop and verify techniques for separating speech that have real benefits for human listeners, both normal and hearing impaired. The PIs have ongoing relationships with several hearing aid manufacturers who will be well paced to bring the benefits of this work to the marketplace.

5 Resources made available

Several resources relating to this project have been made available online at `http://labrosa.ee.columbia.edu/speechsep/`, including examples of real-world speech separation problems, a package of code for aligning recordings to phone labels with very high resolution (developed in order to prepare the tokens for our listening tests), and class materials for the Columbia Speech and Audio Processing and Recognition course (class notes, practicals etc.). We also host the proceedings of the Statistical and Perceptual Audition (SAPA, `http://www.sapa2006.org/`) workshops that are related to this project.

6 Publications

See references [8, 12, 9, 10, 7, 4, 1, 2, 11].

References

[1] M. Athineos and D. P. W. Ellis. Autoregressive modeling of temporal envelopes. *IEEE Tr. Signal Proc.*, 15(11):5237–5245, 2007. URL `http://www.ee.columbia.edu/~dpwe/pubs/AthinE07-fdlp.pdf`.

[2] M. Mandel and D. P. W. Ellis. EM localization and separation using interaural level and phase cues. In *Proc. IEEE Workshop on Apps. of Sig. Proc. to Audio and Acous.*, pages 275–278, Mohonk, NY, 2007. URL `http://www.ee.columbia.edu/~dpwe/pubs/MandE07-ild.pdf`.

[3] H. Nam and L. Goldstein. TADA: TAsk Dynamics Application, 2006. URL `http://www.haskins.yale.edu/tada_download/index.html`. Downloaded 2008-01-29 from `http://www.haskins.yale.edu/tada_download/`.

[4] S. Ravuri and D. P. W. Ellis. Stylization of pitch with syllable-baed linear segments. In *Proc. ICASSP*, Las Vegas, 2008. URL `http://www.ee.columbia.edu/~dpwe/pubs/RavuE08-prosody.pdf`.

[5] S. Rennie, P. Olsen, J. Hershey, and T. Kristjansson. The Iroquois model: Using temporal dynamics to separate speakers. In *Workshop on Statistical and Perceptual Audio Processing (SAPA)*, Pittsburgh, PA, September 2006.

[6] E. Saltzman and K. G. Munhall. A dynamical approach to gestural patterning in speech production. *Ecological Psychology*, 1:333–382, 1989.

[7] Y. Shao, S. Srinivasan, and D.L. Wang. Incorporating auditory feature uncertainties in robust speaker identification. In *Proc. ICASSP*, pages IV–277–280, Hawai'i, Apr 2007.

[8] Barbara G. Shinn-Cunningham and Dali Wang. Influences of auditory object formation on phonemic restoration. *J. Acoust. Soc. Am.*, 123(1):295–301, Jan 2008.

[9] S. Srinivasan and D.L. Wang. Transforming binary uncertainties for robust speech recognition. *IEEE Tr. Audio, Speech, Lang. Proc.*, 15:2130–2140, 2007.

[10] S. Srinivasan, N. Roman, and D.L. Wang. Exploiting uncertainties for binaural speech recognition. In *Proc. ICASSP*, pages IV–789–792, Honolulu, Hawai'i, Apr 2007.

[11] R. J. Weiss and D. P. W. Ellis. Monaural speech separation using source-adapted models. In *Proc. IEEE Workshop on Apps. of Sig. Proc. to Audio and Acous.*, pages 114–117, Mohonk NY, 2007. URL `http://www.ee.columbia.edu/~dpwe/pubs/WeissE07-spkrs.pdf`.

[12] Jin Z. and Wang D.L. A supervised learning approach to monaural segregation of reverberant speech. In *Proc. ICASSP*, pages IV–921–924, Honolulu, Hawai'i, Apr 2007.