Chapter 1

EVALUATING SPEECH SEPARATION SYSTEMS

Daniel P.W. Ellis
LabROSA, Columbia University
New York NY, U.S.A.
dpwe@ee.columbia.edu

Abstract Common evaluation standards are critical to making progress in any field, but they can also distort research by shifting all the attention to a limited subset of the problem. Here, we consider the problem of evaluating algorithms for speech separation and acoustic scene analysis, noting some weaknesses of existing measures, and making some suggestions for future evaluations. We take the position that the most relevant 'ground truth' for sound mixture organization is the set of sources perceived by human listeners, and that best evaluation standards would measure the machine's match to this perception at a level abstracted away from the low-level signal features most often considered in signal processing.

Keywords: speech, separation, evaluation, recognition, CASA, ASR

1. The ASR experience

Quantitative evaluation is an essential and sensitive factor in any area of technological research. Automatic Speech Recognition (ASR) provides an instructive example of the benefits and costs of common evaluation standard. Prior to the mid-1980s, speech recognition research was a confusing and disorganized field, with individual research groups tending to use idiosyncratic measures that showed their particular systems in the best light. Widespread frustration at the difficulty of comparing the achievements of different groups – among researchers and funders alike – was answered by a series of carefully-designed evaluation tasks created by the US National Institute of Standards and Technology (Pallet, 1985). While the speech material in these tasks has evolved from highly constrained vocabularies, grammars and speaking styles through to unconstrained telephone conversations, the principal figure of merit

has remained the Word Error Rate (WER) – the number of incorrect word tokens generated by the system as a percentage of the word count of the ideal transcript – throughout this period.

Over more than 15 years of NIST evaluations, the benefits of the common evaluation task and single performance measure have been dramatic. Standard measures have made it possible to give definitive answers to questions over the relative benefits of different techniques, even when those differences are small. Since the advent of Gaussian Mixture Model-Hidden Markov Model (GMM-HMM) recognition systems, it turns out that most ASR improvements have been incremental, rarely affording an improvement of more than 10% relative, yet we have seen a compound system improvement of perhaps two orders of magnitude through the careful and judicious combination of many small enhancements. Accurate and consistent performance measures are crucial to making this possible.

The disadvantage to this powerful organization of the field around a common goal and metric comes from the kind of 'monoculture' we see in current speech recognition research. Of the many hundreds of papers published in speech recognition journals and conference proceedings each year, the vast majority use same GMM-HMM framework or very close relatives, and of the dozen or so labs working on large-vocabulary speech recognition systems and participating in current NIST evaluations, all are using systems that appear identical to a casual observer. If GMM-HMM systems were obviously the 'right' solution, this might be expected; however, many researchers are uncomfortable with the HMM framework, but feel obliged to keep working with it because the performance loss incurred by switching to a less mature, less optimized novel approach would jeopardize the acceptance (and publication) of their work (Bourlard et al., 1996).

The dominance of a common standard can have other disadvantages. The universal adoption of WER as the principal performance measure has led to a focus on transcription tasks and speech-only material to the neglect of other kinds of signals (including, of particular relevance to the current volume, many kinds of speech-interference mixtures). A single style of task and a single performance measure dominating the field for several decades has resulted in solutions more and more closely optimized for that one task, and a widening gap between performance on the focus tasks and other applications, for instance speech mixtures, that may be equally important in a broad sense but happen not to have been included in the evaluations.

Lessons of evaluation

From the example of speech recognition, we can draw the following lessons:

- Common evaluation tasks (along with the corresponding performance metrics) can have a very positive effect on research and progress in a given field by providing detailed, quantitative answers to questions over the relative merits of different approaches. In addition to furthering debate, this information makes it easier for funding sources to support the field, since they can be more confident that their money is getting results.
- When a single task is defined, and particularly when it bears on funding and other resource allocation, there will be a great concentration on that task leading to the neglect of similar but distinct problems. Thus, the task chosen should ideally represent a real problem with useful applications – so that even in the worst case, with only that one problem being solved, there is still some valuable output from the research.
- Funneling all the effort of a research community into a single, narrow focus is generally undesirable; one alternative is to define more than one task and/or more than one performance measure, to create multiple 'niches' supporting several different threads of research. Making these niches too numerous, however, defeats the benefits of common evaluation: if each separate group evaluates their approach with a different measure, benefits of a common standard are largely lost.

2. Evaluating speech separation

An evaluation task consists of two components: a *domain* or application area, specifying the kinds of material that will be considered (such as in-car speech, or target-versus-interferer speech); a *metric* such as word error rate or signal-to-noise ratio, which implicitly defines the core nature of the problem to be addressed (recognizing spoken words or reducing distortion energy respectively). Since certain metrics place additional constraints on the domain (such as the availability of isolated pre-mixture sources), we will first consider the range of metrics that are available and that have been used in speech separation work.

Metrics can be arranged on an axis of abstraction, from those that measure the most concrete, literal properties of signals, through to those concerned with much higher-level, derived properties in the information extracted from the signals.

Signal-to-noise ratio

The simplest measure, signal-to-noise ratio (SNR), requires that the system being measured reconstructs actual waveforms corresponding to individual sources in a mixture, and that the pre-mixture waveforms of those sources (the 'ideal' outputs) are available. SNR is defined as ratio of the energy of the original target source to the energy of the difference between original and reconstruction – that is, the energy of a signal which, when linearly added to the original, would give the reconstruction. This measure is commonly used for low-level algorithms that have a good chance at near-perfect separation (such as multi-channel Independent Component Analysis (Bell and Sejnowski, 1995), or time-frequency masked reconstruction (Brown and Cooke, 1994)), and is arguably *sufficient*: if we are able to reconstruct a signal that (almost) exactly matches some clean, pre-mixture version, then any other information we wish to obtain is likely also to be available. However, the problems of SNR are:

- It requires the original signal for comparison, largely limiting its use to mixtures that are artificially constructed, rather than those recorded from real environments.
- Distortions such as fixed phase/time delays or nonuniform gains across frequency which can have only a small effect on the perceived quality of a reconstructed sound, can have a large negative effect on SNR.
- The common unit of measurement, energy, has in general only an indirect relationship to perceived quality. The same amount of energy will have a widely-varying impact on perceived quality depending on where and how it is placed in time-frequency; this is particularly significant in the case of speech, where *most* of the energy is below 500 Hz, yet very little intelligibility is lost when this energy is filtered out. Another example of the disconnect between SNR and perceived quality comes from the psychoacoustic-based coding used in schemes like 'MP3' audio, where a reproduction with an SNR of under 20 dB can sound essentially perfect because all the distortion energy has been carefully hidden below the complex masking patterns of the auditory system.

Representation-based metrics

While SNR has the attraction of being applicable to any system that generates an output waveform, more helpful measures (at least from the point of view of system development) can be derived directly from

whatever representation is used within a particular system. Thus, in Cooke's original Computational Auditory Scene Analysis (CASA) system (Cooke, 1991), an evaluation was performed by comparing the 'strands' representations resolved by his system with the representation generated for each source in isolation, thereby avoiding the need for a strands-to-sound resynthesis path.

By considering the internal representation, evaluations can also be made relative to an 'ideal' performance that reflects intrinsic limitations of a given approach. Many recent systems are based on time-frequency (TF) masked refiltering, in which Gabor 'tiles' in TF are classified as target-dominated and selectively resynthesized (Hu and Wang, 2003; Roweis, 2001). Such an approach cannot separate overlapped energy falling into a single cell, so an SNR ceiling is achieved by an 'ideal' mask consisting of all the cells in which target energy is greater than interference (since including any other cells will increase the distortion energy, by adding noise, more than it is decreased by reducing the deleted target). Systems based on these masks can be evaluated by how closely they approach this ideal mask e.g. by measuring the classification accuracy of TF cells. This measure removes the effective weighting of each cell by its local signal energy in SNR calculation; however, it gives a disproportionate influence to near-silent TF cells whose 'ideal' classification will depend on the likely irrelevant noise-floor level in the original mixture components.

Another analysis possible with masked refiltering systems is the separate accounting for distortion due to energy deleted from the target, and due to included portions of the interference (the "energy loss" and "noise residue" of (Hu and Wang, 2003)). However, we are again faced by the problem of the perceptual incomparability of energy in different parts of time-frequency.

Perceptual Models

As indicated above, the inadequacies of SNR have long been apparent in the active and successful field of audio coding. When the goal is to satisfy human listeners (e.g. telephony customers or music consumers), there is no substitute for formal listening tests in which subjects rate the perceived quality of various algorithms applied to the same material. Due to the cost of such evaluations, however, considerable effort has gone into developing algorithmic estimates such as the ITU standards for PEAQ and PESQ (Perceptual Evaluation of Audio/Speech Quality (Thiede et al., 2000)). While these measures also require a predistortion original reference, they make sophisticated efforts to factor out

perceptually-irrelevant modifications, and their use for the evaluation of low-level signal-separation systems deserves investigation.

High-level attributes

While perfect signal recovery may be sufficient, it is rarely necessary. Signal separation is not an end in itself, but a means to some subsequent application, be that recognizing the words in some noisy speech, or even the pleasure of listening to a solo musical performance without additional instruments. In every case, metrics can be devised to measure more directly the success of the signal separation stage on the overall application. When the ultimate task is extracting specific parameters, such as the times of occurrence of certain events, or perhaps limited descriptions of such events (such as onset times, pitches, and intensities in polyphonic music transcription), it is natural to evaluate in terms of the error in that domain.

By far the most widespread evaluation falling into this category is word error rate of speech recognition systems for mixed signals. Given the widespread acceptance of WER as a measure for isolated speech recognition, it is natural to extend the same metric to conditions of significant interference, even when substantially different processing is introduced to address that interference. This approach was taken in one of the earliest models of auditory scene analysis (Weintraub, 1985), although in that work, as in many subsequent experiments, it was found that in some cases that the signal separation preprocessing made the error rate worse than simply feeding the original mixture to an unmodified recognition engine.

Using signal separation to aid speech recognition requires a careful match between the separation techniques and the recognition engine: one example of a well-matched combination highlights the difference between this and lower-level metrics. In missing-data recognition (Cooke et al., 2001), the matching between observed signal and learned speech models is modified to account for limited observability of the target i.e. that certain dimensions can be missing at different times. Acoustic scene organization algorithms are then employed to indicate which dimensions (e.g. TF cells) are reliable correlates of the target speech at each instant (Barker et al., 2004). This work reports minimal increase in word error rate in cases when significantly less than half the features are deemed 'reliable' – a situation which would likely give a highly-distorted resynthesis, but which still contains plenty of information to recognize the spoken words largely without ambiguity.

However, because of the sensitivity of WER measures to the compatibility (and, as in (Barker et al., 2004), close functional integration) between separation algorithm and speech recognizer, this measure is only appropriate for systems specifically built for this application.

Domains

Among acoustic signal separation tasks, speech mixed with different kinds of interference is the most popular domain and is our main concern here. The target voice can experience different amounts of spectral coloration, reverberant smearing, or other distortion, but the main axis of variation is in the nature of the interference signal. Simplest is Gaussian white noise, which can be made a more relevant masker by filtering into pink noise (equally energy per octave) or to match some average speech spectrum. In speech recognition, a common approach is to train models for the combination of speech-plus-noise, which can be very successful for such stationary noise particularly when the absolute level is constrained; a distinct stage of signal separation is avoided completely.

Real-world sound sources comprise a more challenging form of interference because their impact on the features cannot be predicted so accurately (i.e. with small variance), although the combination of a large number of independent sources will tend towards Gaussian noise. At a given power level, the most difficult interference should be a single second voice, since the statistical features of the interference are indistinguishable from the target. (In practice, a single voice offers many opportunities 'glimpsing' the target during silent gaps in the interference, so a combination of a small number of unsynchronized voices may achieve greater interference.)

The majority of noisy speech tasks are created artificially by mixing speech recorded in quiet conditions with different "pure interference" signals (e.g. (Pearce and Hirsch, 2000)). This approach has the attractions that the relative levels of speech and noise can be adjusted, the same speech can be embedded in several different types of noise, and the clean speech can be used to generate a baseline performance. However, it is a poor match to reality: there is no guarantee that the synthetic mixture actually resembles something that could ever have been recorded in a noisy environment, not least because of the Lombard effect, the reflexive modification of speaking quality employed by humans to overcome noisy environments (Lane and Tranel, 1971). Other effects such as reverberation are also frequently ignored in synthetic noise mixtures.

Given the problem identified above of solving only what we test, it would seem preferable to use real recordings of speech in noisy environ-

ments as test material. While some data of this kind do exist (Schmidt-Nielsen et al., 2000), on the whole it is avoided due to the flexibility and control available with synthetic mixtures as just mentioned: recording a range of speech material against a range of background noise types and levels requires, in the worst case, a separate recording for each condition, rather than factorized combinations of a few base recordings.

Another problem with real-world recordings is the availability of ground-truth descriptions. If we artificially mix a clean voice signal against a noisy background, we may hope that our speech separation algorithms will recreate the original clean speech; if the noisy speech is all we have, how can we even judge the quality of the resynthesis? I would argue, however, that this assumption that the pre-mixture original represents the unique best output we could hope for is in fact dodging the more difficult, but more important, question of deciding what we *really* want. If the purpose of the algorithm is to enhance noisy speech for a person to listen to, then the appropriate metric is subjective quality rating, not similarity to an original signal which may not, in fact, match the impression of the source speech in the mind of the listener.

This appeal to subjective sources for ground truth in complex mixtures extends beyond speech in noise: In (Ellis, 1996), a computational auditory scene analysis system that sought to mark the occurrence of different sound events in complex, real-world ambient sounds was evaluated on its ability to duplicate the consensus results of a set of listeners given the same task.

3. Conclusions and recommendations

In light of this discussion, we make some recommendations for the form of a future, widely-applicable evaluation task for speech separation systems:

- It should be based on some kind of real-world task, so that if the worst-case occurs and we end up with solutions applicable only to this narrow task, they can at least be deployed to some purpose.

- The data should be real recordings, or possibly synthetic recordings in which all the possibly relevant aspects of the real recording have been carefully duplicated.

- The evaluation ground truth (be it word transcripts, event detection and descriptions, or other information from the signal) should originate from human transcribers to get at the 'subjective' character of the sound.

- As this implies, the domain of comparison should be in terms of high-level information and attributes, rather than low-level comparisons against some ideal waveform.
- If the task represents a real and useful domain, it ought to be possible to gather comparable human performance on the same task, so we can accurately measure how well our machines do relative to the best currently-known listening machine. Ideally, this would be a task that humans (perhaps impaired populations) find somewhat difficult, to give the machines a chance to exceed human performance – although machines that came anywhere close to human performance on any kind of acoustic scene analysis would be welcome.

One possible domain is audio recorded in real, multi-party meetings, and this task has recently begun to attract attention (Yu et al., 1999; Morgan et al., 2001). Such corpora typically involve significant amounts of speech overlap, and often have both near- and far-field microphone recordings; the head-mounted near-field mics provide a kind of ground-truth reference for the voices picked up by the far-field tabletop mics.

Speech separation is often referred to as the Cocktail-Party problem (following (Cherry, 1953)), and a room containing multiple simultaneous conversations might provide an interesting test domain, one that would mostly defeat human listeners. Such a party could be staged with each participant wearing a head-mounted microphone (which can be inconspicuous) to provide some level of ground-truth. An interesting corpus along these lines is the Sheffield-ATR Crossword task (Crawford et al., 1994), which involved two simultaneous conversations with a fifth participant occasionally involved in both.

A final area is the kind of continuous personal recording proposed in (Bush, 1945) and investigated in (Clarkson et al., 1998): wearable microphones and miniature hard-disk recorders can easily make complete records of a user's acoustic environment, but to allow any kind of useful retrieval from hundreds of hours of such recordings requires automatic analysis of which acoustic source separation will be an important part.

In conclusion, insights from speech recognition and elsewhere show that a common evaluation task is critical to the future progress and support of speech separation research. The form and nature of such a task, however, is far from clear, not least because there is little consensus on the real purpose or ultimate application for speech separation technologies. We favor a task firmly embedded in real-world scenario, and an evaluation metric that reflects subjective information extraction rather than an objective, but arbitrary, low-level ideal.

References

Barker, Jon, Cooke, Martin, and Ellis, Dan (2004). Decoding speech in the presence of other sources. Submitted to Speech Communication.

Bell, Anthony J. and Sejnowski, Terrence J. (1995). An information-maximization approach to blind separation and blind deconvolution. *Neural Computation*, 7(6):1129–1159.

Bourlard, H., Hermansky, H., and Morgan, N. (1996). Towards increasing speech recognition error rates. *Speech Communication*, pages 205–231.

Brown, G. J. and Cooke, M. (1994). Computational auditory scene analysis. *Computer speech and language*, 8:297–336.

Bush, V. (1945). As we may think. *The Atlantic Monthly*.

Cherry, E.C. (1953). Some experiments on the recognition of speech with one and two ears. *J. Acoust. Soc. Am.*, 25:975–979.

Clarkson, B., Sawhney, N., and Pentland, A. (1998). Auditory context awareness via wearable computing. In *Proc. Perceptual User Interfaces Workshop*.

Cooke, M. P. (1991). *Modelling auditory processing and organisation*. PhD thesis, Department of Computer Science, University of Sheffield.

Cooke, Martin, Green, Phil, Josifovski, Lubomir, and Vizinho, Ascencion (2001). Robust automatic speech recognition with missing and unreliable acoustic data. *Speech Communication*, 34(3):267–285.

Crawford, M.D., Brown, G.J., Cooke, M.P., and Green, P.D. (1994). Design, collection and analysis of a multi-simultaneous-speaker corpus. In *Proc. Inst. Acoustics*, volume 5, pages 183–190.

Ellis, D. P. W. (1996). *Prediction–driven computational auditory scene analysis*. PhD thesis, Department of Electrtical Engineering and Computer Science, M.I.T.

Hu, G. and Wang, D.L. (2003). Monaural speech separation. In *Advances in NIPS 13*, Cambridge MA. MIT Press.

Lane, H. and Tranel, B. (1971). The Lombard sign and the role of hearing in speech. *J. Speech and Hearing Res.*, (14):677–709.

Morgan, N., Baron, D., Edwards, J., Ellis, D., Gelbart, D., Janin, A., Pfau, T., Shriberg, E., and Stolcke, A. (2001). The meeting project at ICSI. In *Proc. Human Lang. Tech. Conf.*, pages 246–252.

Pallet, D.S. (1985). Performance assessment of automatic speech recognizers. *J. Res. Natl. Bureau of Standards*, 90:371–387.

Pearce, D. and Hirsch, H.-G. (2000). The AURORA experimental framework for the performance evaluation of speech recognition systems under noisy conditions. In *Proc. ICSLP '00*, volume 4, pages 29–32, Beijing, China.

Roweis, S. (2001). One-microphone source separation. In *Advances in NIPS 11*, pages 609–616. MIT Press, Cambridge MA.

Schmidt-Nielsen, Astrid, Marsh, Elaine, Tardelli, John, Gatewood, Paul, Kreamer, Elizabeth, Tremain, Thomas, Cieri, Christopher, and Wright, Jon (2000). *Speech in Noisy Environments (SPINE) Evaluation Audio*. Linguistic Data Consortium, Philadelphia PA.

Thiede, T., Treurniet, W.C., Bitto, R., Schmidmer, C., Sporer, T., Beerends, J.G., Colomes, C., Keyhl, M., Stoll, G., Brandeburg, K., and Feiten, B. (2000). PEAQ – the ITU standard for objective measurement of perceived audio quality. *J. Audio Eng. Soc.*, 48(1/2).

Weintraub, M. (1985). *A theory and computational model of auditory monoaural sound separation*. PhD thesis, Department of Electrical Engineering, Stanford University.

Yu, H., Finke, M., and Waibel, A. (1999). Progress in automatic meeting transcription.