1.3 Evaluation

As mentioned in Sec. 1, the common goal of many speech enhancement, and multi-talker speech separation, algorithms is to improve either speech quality or intelligibility, of a degraded speech signal. But how do you accurately evaluate if an algorithm-under-test really does improve one of these quantities?

In general, the only way to truly evaluate if a speech processing algorithm in fact does improve speech quality or intelligibility, is by a listening test involving the end user, i.e. human test subjects. However, listening tests are involved as they require numerous human test subjects and the listening test itself, needs to be carefully planned based on whether the goal is to evaluate speech intelligibility or speech quality. Most people probably have an idea about what a good quality speech signal sounds like, and what would make the same signal a bad quality one, e.g. by introducing hiss or crackle sounds to the signal. Nevertheless, speech quality is highly subjective as it is primarily based on emotions and feelings. Speech intelligibility, on the other hand, is much more objective, if you will, as emotions and feelings in general do not influence your capability of understanding speech. Either you understand what is being said or you do not. Consequently, designing listening tests that truly evaluate speech quality or intelligibility, is no easy task [21, 22].

Therefore, to avoid these often tedious and time consuming listening tests, and to get a quick and somewhat accurate estimate of the listening-test result, a set of objective measures have been designed, which are based on mathematical functions that quantify the difference between clean and noisy/processed speech signals in a way that has a high correlation with listening-test results. In fact, in some cases, it is more desirable to use an objective measure, instead of a listening test involving human test subjects, as objective measures are fast, cheap, and consistently produce the same result for the same testing condition, whereas listening-test results might vary due to factors such as listener fatigue, or varying hearing ability among test subjects.

In the following, three of the popular techniques for objective quality and intelligibility evaluation are briefly reviewed.

1.3.1 Perceptual Evaluation of Speech Quality

The Perceptual Evaluation of Speech Quality (PESQ) [121–124] measure is one of the most widely used objective measures for estimating speech quality [22]. The PESQ measure is designed to approximate the Mean Opinion Score (MOS), which is a widely used listening test procedure for speech quality evaluation [21, 22, 122, 125]. The MOS is a very simple evaluation procedure, where the test subjects are asked to grade the speech signal they are hearing based on a scale with five discrete steps, with "1" representing a bad and very annoying sound quality, and "5" representing an excellent sound quality with imperceptible distortions. The final MOS score, which is a single scalar between "1" and "5", is simply the average, or mean, of all the "opinion scores" for each test signal and for all test subjects, hence the name, mean opinion score. As mentioned, the PESQ measure approximates MOS, but the PESQ algorithm is fairly complex as it consists of multiple steps involving pre-processing, time alignment, perceptual filtering, masking effects, etc. (see e.g. [22, pp. 491-503]). Nevertheless, PESQ versions P.862.1/2 [123, 124] produce a number ranging from approximately 1 to 4.5, which allow comparisons between PESQ and MOS, and PESQ has been found to be highly correlated with listening-test experiments based on MOS [121, 123]. In fact, although PESQ was originally designed for evaluating speech coding algorithms, it was later shown that PESQ correlated reasonably well with the quality of speech processed by commonly used speech enhancement algorithms [126]. Also, PESQ requires both the clean speech signal as well as the noisy/processed signal to estimate the perceived quality of the noisy/processed signal. This makes PESQ an intrusive speech quality estimator, which limits its use to situations where the clean undistorted signal is available in isolation. For most applications of PESQ, this is not a real limitation as PESQ is usually used in laboratory conditions, where the clean signal is often available in isolation.

1.3.2 Short-Time Objective Intelligibility

The Short-Time Objective Intelligibility (STOI) [127, 128] is, today, perhaps, the most widely used objective measure for estimating speech intelligibility. Differently from PESQ, STOI is not designed to approximate any specific type of listening test, but merely designed to correlate well with listening tests evaluating speech intelligibility in general. Since intelligibility is binary in the sense that, either a given speech signal, say a word, is understood or it is not, listening-test results representing speech intelligibility can, most often, be quantified as a number between 0 and 1 that represents the percentage of words correctly understood [22]. To be comparable with such tests, STOI is designed to produce a single scalar output in a similar range[5], with an output of 1 indicating fully intelligible speech.

Similarly to PESQ, STOI is an intrusive algorithm as it requires both the clean signal and the noise/processed signal in isolation. Furthermore, STOI is based on the assumption that modulation frequencies play an important role in speech intelligibility, and that all frequency bands in the cochlear filter are equally important. These are assumptions, which, to a certain degree,

[5] In theory, STOI can produce numbers in the interval $(-1, 1)$, since STOI is based on a correlation coefficient measure. However, in practice, negative numbers are rarely observed.

are justified empirically [4, 129, 130]. This also has the consequence that, compared to PESQ, STOI is a fairly simple algorithm.

Despite its simple formulation, STOI has been found to be able to quite accurately predict the intelligibility of noisy/processed speech in a wide range of acoustic scenarios [128, 131–134]. Finally, an extension to STOI, known as Extended Short-Time Objective Intelligibility (ESTOI), has been proposed as a more accurate speech intelligibility predictor in the special cases where the noise sources are highly modulated [135].

1.3.3 Blind Source Separation Evaluation

When evaluating single-target signal speech processing algorithms, such as speech enhancement, PESQ and STOI are useful, as these measures quantify how successful the algorithm-under-test process a degraded signal in a way that is perceptually desirable. If, however, multiple target signals exist, such as in a speech separation task, additional information about the processing artifacts might be desirable compared to what PESQ and STOI can provide [136, 137]. In other words, when a mixture signal that contains multiple speech and noise signals are processed by a speech separation algorithm, the enhanced or separated speakers might contain artifacts originating from multiple different sources. For example, these artifacts could originate from the noise signal itself, from processing artifacts, or due to "cross-talk", i.e. signal components from one target speaker appearing in the separated signal of the other.

One of the most popular objective measures for evaluating speech separation algorithms that take these considerations into account, is the Blind-Source Separation (BSS) Eval toolkit [138]. In the technique proposed in [138], the separated signals are decomposed into target-speaker components and three noise components known as interference, noise, and artifact. The interference component represents cross-talk from other target speakers. Noise and artifacts, represents environmental noise sources and processing artifacts, respectively. From this decomposition, energy-ratio measures are defined known as Source-to-Distortion Ratio (SDR), Source-to-Interference Ratio (SIR), Source-to-Artifact Ratio (SAR), and SNR, which each relate these decomposed elements of the separated signal in a way that provide useful information about the contribution of each of them. Finally, it has been found that these objective measures correlate well with listening test evaluating quality [139, 140], and, obviously, the BSS Eval toolkit only compliments other objective measures such as STOI and PESQ.

[112] S. T. Roweis, "One Microphone Source Separation," in *Advances in Neural Information Processing Systems 13*, T. K. Leen, T. G. Dietterich, and V. Tresp, Eds. MIT Press, 2001, pp. 793–799.

[113] A. Ozerov, C. Févotte, and M. Charbit, "Factorial Scaled Hidden Markov Model for polyphonic audio representation and source separation," in *Proc. WASPAA*, 2009, pp. 121–124.

[114] T. Virtanen, "Speech Recognition Using Factorial Hidden Markov Models for Separation in the Feature Space," in *Proc. INTERSPEECH*, 2006, pp. 89–92.

[115] J. R. Hershey and M. Casey, "Audio-Visual Sound Separation Via Hidden Markov Models," in *Proc. NIPS*, 2002, pp. 1173–1180.

[116] M. Stark, M. Wohlmayr, and F. Pernkopf, "Source-Filter-Based Single-Channel Speech Separation Using Pitch Information," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 2, pp. 242–255, 2011.

[117] G. J. Mysore, P. Smaragdis, and B. Raj, "Non-negative Hidden Markov Modeling of Audio with Application to Source Separation," in *Latent Variable Analysis and Signal Separation*, ser. Lecture Notes in Computer Science. Springer, 2010, pp. 140–148.

[118] T. T. Kristjansson, J. R. Hershey, P. A. Olsen, S. J. Rennie, and R. A. Gopinath, "Super-human multi-talker speech recognition: the IBM 2006 speech separation challenge system," in *Proc. INTERSPEECH*, 2006, pp. 97–100.

[119] J. R. Hershey, S. J. Rennie, P. A. Olsen, and T. T. Kristjansson, "Super-human multi-talker speech recognition: A graphical modeling approach," *Computer Speech & Language*, vol. 24, no. 1, pp. 45–66, 2010.

[120] Y.-m. Qian, C. Weng, X.-k. Chang, S. Wang, and D. Yu, "Past review, current progress, and challenges ahead on the cocktail party problem," *Frontiers of Information Technology & Electronic Engineering*, vol. 19, no. 1, pp. 40–63, 2018.

[121] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs," in *Proc. ICASSP*, vol. 2, 2001, pp. 749–752.

[122] "International Telecommunication Union - Recommendation BS.562 : Subjective assessment of sound quality," 1990.

[123] "International Telecommunication Union - Recommendation P.862.1 : Mapping function for transforming P.862 raw result scores to MOS-LQO," 2003.

[124] "International Telecommunication Union - Recommendation P.862.2 : Wideband extension to Recommendation P.862 for the assessment of wideband telephone networks and speech codecs." 2005.

[125] IEEE, "IEEE Recommended Practice for Speech Quality Measurements," *IEEE No 297-1969*, pp. 1–24, 1969.

[126] Y. Hu and P. C. Loizou, "Evaluation of Objective Quality Measures for Speech Enhancement," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 1, pp. 229–238, 2008.

[127] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "A short-time objective intelligibility measure for time-frequency weighted noisy speech," in *Proc. ICASSP*, 2010, pp. 4214–4217.

[128] ——, "An Algorithm for Intelligibility Prediction of Time-Frequency Weighted Noisy Speech," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2125–2136, 2011.

[129] T. M. Elliott and F. E. Theunissen, "The Modulation Transfer Function for Speech Intelligibility," *PLOS Computational Biology*, vol. 5, no. 3, 2009.

[130] R. Drullman, J. M. Festen, and R. Plomp, "Effect of temporal envelope smearing on speech reception," *The Journal of the Acoustical Society of America*, vol. 95, no. 2, pp. 1053–1064, 1994.

[131] S. Jørgensen, J. Cubick, and T. Dau, "Speech Intelligibility Evaluation for Mobile Phones." *Acustica United with Acta Acustica*, vol. 101, pp. 1016–1025, 2015.

[132] T. H. Falk *et al.*, "Objective Quality and Intelligibility Prediction for Users of Assistive Listening Devices: Advantages and limitations of existing tools," *IEEE Signal Processing Magazine*, vol. 32, no. 2, pp. 114–124, 2015.

[133] J. Jensen and C. H. Taal, "Speech Intelligibility Prediction Based on Mutual Information," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 2, pp. 430–440, 2014.

[134] R. Xia, J. Li, M. Akagi, and Y. Yan, "Evaluation of objective intelligibility prediction measures for noise-reduced signals in mandarin," in *Proc. ICASSP*, 2012, pp. 4465–4468.

[135] J. Jensen and C. H. Taal, "An Algorithm for Predicting the Intelligibility of Speech Masked by Modulated Noise Maskers," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 11, pp. 2009–2022, 2016.

[136] E. Vincent, H. Sawada, P. Bofill, S. Makino, and J. P. Rosca, "First Stereo Audio Source Separation Evaluation Campaign: Data, Algorithms and Results," in *Independent Component Analysis and Signal Separation*, ser. Lecture Notes in Computer Science. Springer, 2007, pp. 552–559.

[137] V. Emiya, E. Vincent, N. Harlander, and V. Hohmann, "Subjective and Objective Quality Assessment of Audio Source Separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2046–2057, 2011.

[138] E. Vincent, R. Gribonval, and C. Fevotte, "Performance measurement in blind audio source separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 4, pp. 1462–1469, 2006.

[139] D. Ward, H. Wierstorf, R. Mason, E. Grais, and M. Plumbley, "BSS eval or PEASS? Predicting the perception of singing-voice separation," in *Proc. ICASSP*, 2018, pp. 596 – 600.

[140] E. Cano, D. FitzGerald, and K. Brandenburg, "Evaluation of quality of sound source separation algorithms: Human perception vs quantitative metrics," in *Proc. EUSIPCO*, 2016, pp. 1758–1762.

[141] W. Schultz, P. Dayan, and P. R. Montague, "A Neural Substrate of Prediction and Reward," *Science*, vol. 275, no. 5306, pp. 1593–1599, 1997.