Can we Automatically Transform Speech Recorded on Common Consumer Devices in Real-World Environments into Professional Production Quality Speech? — A Dataset, Insights, and Challenges

Gautham J. Mysore, *Member, IEEE,*

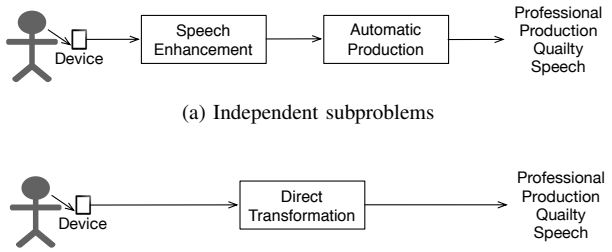
***Abstract*—The goal of speech enhancement is typically to recover clean speech from noisy, reverberant, and often bandlimited speech in order to yield improved intelligibility, clarity, or automatic speech recognition performance. However, the acoustic goal for a great deal of speech content such as voice overs, podcasts, demo videos, lecture videos, and audio stories is often not merely clean speech, but speech that is aesthetically pleasing. This is achieved in professional recording studios by having a skilled sound engineer record clean speech in an acoustically treated room and then edit and process it with audio effects (which we refer to as production). A growing amount of speech content is being recorded on common consumer devices such as tablets, smartphones, and laptops. Moreover, it is typically recorded in common but non-acoustically treated environments such as homes and offices. We argue that the goal of enhancing such recordings should not only be to make it sound cleaner as would be done using traditional speech enhancement techniques, but to make it sound like it was recorded and produced in a professional recording studio. In this paper, we show why this can be beneficial, describe a new data set (a great deal of which was recorded in a professional recording studio) that we prepared to help in developing algorithms for this purpose, and discuss some insights and challenges associated with this problem.**

***Index Terms*—Speech Enhancement, Automatic Production.**

I. Introduction

LARGE amounts of speech content such as voice overs, podcasts, demo videos, lecture videos, and audio stories are regularly recorded in non-professional acoustic environments such as in homes and offices. Moreover, this is often done with common consumer devices such as tablets, smartphones, and laptops. Although these recordings are typically intelligible, they often sound of poor quality, and it is generally apparent that they were not professionally created. Some reasons for this are that they suffer from ambient noise, reverberation, low quality and often bandlimited recording hardware (microphone, microphone preamplifier, and analog to digital converter on a device), and have not been professionally produced. We refer to these recordings as device recordings.

When such content is created in a professional recording studio, a skilled sound engineer typically performs a clean recording in an acoustically treated low noise, low reflection vocal booth with high quality recording equipment [1]. The sound engineer then removes non-speech sounds such as breaths and lip smacks, and finally applies audio effects such as an equalizer, dynamic range compressor, and de-esser to make it sound more aesthetically pleasing (production) [2]–[4]. We refer to these recordings as produced recordings.



(a) Independent subproblems

(b) Single problem

Fig. 1. One could attempt to solve the problem by treating it as two independent subproblems with the intermediate goal of recovering clean speech or as a single problem of directly transforming the device recording.

We argue that if the creator of the kinds of speech content mentioned above had no time or budget restrictions, he or she is likely to create the content in a professional recording studio with the help of a professional sound engineer. However, due to these restrictions a large amount of content is created on common consumer devices. Higher quality microphones and recording equipment are sometimes connected to such devices, but they are still prone to the same ambient noise, reverberation, and lack of production as standard device recordings. Therefore, we believe that it would be highly beneficial to develop algorithms to automatically transform device recordings into produced recordings.

One approach to address this problem is to decompose it into two subproblems — recover clean speech and then perform automatic production on the recovered clean speech estimate (Fig. 1a). Current speech enhancement algorithms address the first subproblem largely by denoising [5]–[8], dereverberation [9], [10], decoloration [11], and to some degree, bandwidth expansion [12], [13] with the goal to improve intelligibility, clarity, or automatic speech recognition performance. A naive approach to the second subproblem is simply to use preset parameter values of audio effects.

Gautham J. Mysore is with Adobe Research, San Francisco, CA 94103, USA (e-mail: gmysore@adobe.com)

However, professional sound engineers carefully listen to the speech content at hand and set the parameters of the effects to sound the best for that content. It would therefore be beneficial for an algorithm to adaptively do this [14], [15].

Given that there is a great deal of existing literature in speech enhancement and some literature in automatic production as mentioned above, one could potentially make use of parts of existing techniques to solve the subproblems. However, it could be beneficial to do so in a way in which the solutions to the subproblems are not completely independent (for reasons described in Section III).

Another approach to address this problem is to directly attempt to transform device speech into produced speech without the intermediate recovery of clean speech (Fig. 1b). In Section III, we show why this could be beneficial. One example of such a transformation could come from a learned non-linear mapping of short time segments of some representation of device speech to that of produced speech using classes of techniques such as deep learning [16] or Gaussian process regression [17].

In order to facilitate research on this problem, we developed the DAPS (device and produced speech) dataset, which is a new, easily extensible dataset (described in Section II) of aligned versions of clean speech, produced speech, and a number of versions of device speech (recorded with different devices in a number of real-world acoustic environments). Additionally, in the accompanying website[1], we outline a procedure for researchers to easily create new versions of device recordings and provide tools to assist in this process. The dataset could also be useful for research in traditional speech enhancement, automatic production of studio recordings, and problems such as voice conversion.

In Section IV, we discuss some of the challenges in evaluation of algorithms to solve this problem, and discuss some potential approaches to evaluation.

II. Dataset

We developed the DAPS (device and produced speech) dataset [1] to facilitate research on transforming device recordings into produced recordings. A major goal in creating this dataset is to provide multiple, aligned versions of speech such that they correspond to real-world examples of inputs and outputs of each block in Fig. 1. They can therefore be used as training data when developing algorithms for this purpose. We describe the different versions below (illustrated in Fig. 2). The first three versions correspond to the standard recording and production pipeline in a professional recording studio. The dataset consists of twenty speakers (ten female and ten male) reading five excerpts each from public domain stories, which yields about fourteen minutes of data per speaker. Each version described below contains all excerpts read by all speakers.

A. Clean Raw

These recordings were performed in an acoustically treated low noise, low reflection vocal booth of a professional recording studio using a microphone with a flat frequency response (Sennheiser MKH 40). In order to create a near anechoic room, a thick curtain was placed in the vocal booth in front of the glass that separates it from the control room. A sampling rate of 88.2 KHz was used for the initial recording (as the use of high sampling rates is now common practice in recording studios), but we provide downsampled versions at 44.1KHz in the dataset. These recordings contain speech as well as some non-speech vocal sounds such as breaths and lip smacks. All other versions are derived from this version.

[1]Available at https://ccrma.stanford.edu/~gautham/Site/daps.html

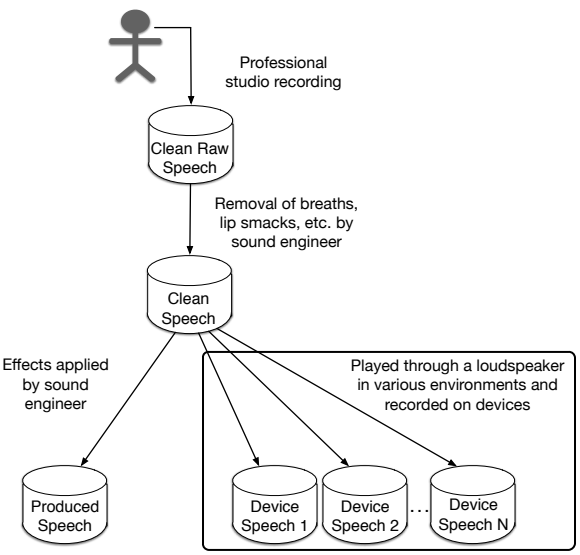


Fig. 2. Illustration of the creation of the DAPS dataset showing the various versions of aligned speech that it includes.

B. Clean

The sound engineer carefully removed most non-speech sounds such as breaths and lip smacks from the clean raw recordings to create this version.

C. Produced

For this version, we asked the sound engineer to perform any processing that he would typically perform in order to make the recordings sound aesthetically pleasing and professionally produced. The only restriction that we placed is that he must use the same effects in the same order for all recordings. He used the following effects from the Izotope Nectar suite of plugins for this purpose in the following order — tape saturation simulator, equalizer, dynamic range compressor, de-esser, limiter. The parameter settings of these effects were different for each speaker and based on what the sound engineer thought sounded the best for a given speaker (but constant for all excerpts of a given speaker).

D. Device

This set of versions correspond to people talking into commonly used consumer devices in real-world acoustic environments. One way to obtain such data is to have them physically perform these recordings in a number of different rooms using different devices. The problems with this approach are that

Fig. 3. Setup for a device recording in a conference room. The clean studio recording is played through the loudspeaker and recorded on a tablet (iPad Air), capturing the noise and reverberation of the room as well as the limitations of the recording hardware.

there will be differences in the speech performance in each room, the device versions will not be perfectly aligned with the studio versions, and the process will be quite laborious when recording multiple versions.

To get around these consistency and labor intensive issues, we could take a more typical approach [18]–[20] used in creating speech enhancement datasets, which is to convolve clean speech with a room impulse response and/or artificially mix it with ambient noise. This has the advantage of the availability of ground truth clean speech data. However, the synthetic nature of the data is not likely to capture all the nuances of a real-world degraded recording.

In an attempt to capture these real-world nuances as well as to provide ground truth data, we took a different approach. For each acoustic environment, we placed a high quality loudspeaker on a table such that the speaker cones are at about the height of a person sitting in a chair in that environment, played the clean version of the recorded speech through the loudspeaker, and recorded it into a device (one instance is shown in Fig. 3). We used a coaxial loudspeaker with built in amplifier (Presonus Sceptre S6 studio monitor) so that it better approximates a point source than a two-way or three-way loudspeaker, and placed it on a stand that decouples vibrations between the loudspeaker and the table. The distance between the loudspeaker and device was about eighteen inches. Speech was played at a typical conversational level.

One design decision was if we should play the clean raw or clean version through the loudspeaker. In other words, the question is if we should leave non-speech vocal sounds such as breaths and lip smacks in the device recordings or not. We chose to play the clean version (without non-speech sounds) so that the only difference between the device recordings and the produced recordings are acoustic qualities. This is likely to help in the development of certain algorithms that attempt to learn a mapping between the device and produced recordings. One could then treat the removal of non-speech vocal sounds as a pre-processing step and use the clean and clean raw data as examples of input and output data for that purpose. Moreover, it is quite simple to create new device versions with the clean raw version as input if desired (as discussed below).

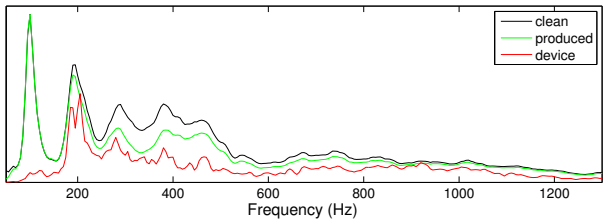


Fig. 4. Average magnitude spectrum of different versions of a given script spoken by a male speaker (zoomed in to a limited frequency range) is an indication of coloration.

Another decision was the choice of devices and acoustic environments for device recordings included in the dataset. We provide twelve versions of device recordings with a tablet (iPad Air) and smartphone (iPhone 5S) in different acoustic environments. In most of the recordings, the device is placed on a stand to simulate a person holding it, but in a few recordings, it is placed flat on a table as this is sometimes the way in which people record on such devices.

The primary goal of creating this dataset was to transform device recordings of the kind of speech content mentioned in Section I into professionally produced versions. Such content is typically recorded in rooms with poor acoustics, a relatively high signal to noise ratio, and relatively stationary noise, so we primarily used such rooms. Specifically we used offices, conference rooms, a living room, and a bedroom. In order to provide a single more challenging acoustic environment, we also used a balcony near a road with heavy traffic.

We used a sampling rate of 44.1 KHz on the device recordings so that they could be aligned to the studio versions. These devices each have multiple microphones, so one can conjecture that some form of multi-channel speech enhancement is performed on the devices. This would mean that the device recordings in this dataset might have undergone some pre-processing. Regardless, this would be the input to an application that one might develop for one of these devices, so we believe that it is the right data to use for this purpose.

We also provide instructions (in the accompanying website) and tools (available with the dataset) to make it simple for researchers to create new device recordings with different devices or microphones in different acoustic environments.

III. Synergy between Subproblems

Since the goal is to obtain produced speech given device speech, rather than to recover intermediate clean speech, one can take advantage of the relationship between certain aspects of the two subproblems (speech enhancement and automatic production). Additionally, when developing algorithms for this purpose, it would be useful to account for certain issues that would not have been present if the goal was to solve a single subproblem. In this section, we highlight a few examples of this synergy between subproblems.

A. Decoloration

Device recordings often have a great deal of coloration with respect to clean recordings due to factors such as the

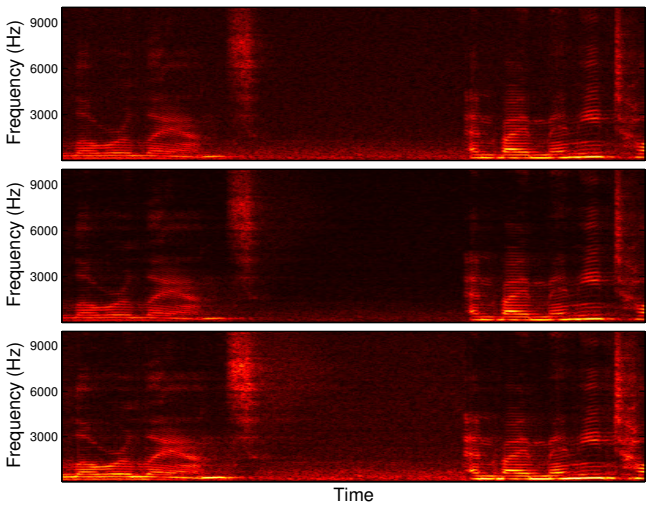


Fig. 5. A clip of a device recording (top) with denoising applied (middle), and a dynamic range compressor applied after denoising (bottom). As shown, dynamic range compression brings the noise floor back up.

short term effects of reverberation and low quality bandlimited recording hardware (Fig. 4). A speech enhancement algorithm would directly [11] or indirectly [8]–[10] apply some form of decoloration and perhaps bandwidth expansion [12], [13]. However, certain effects typically used by a sound engineer, such as an equalizer, also impart coloration. As shown in Fig. 4, although the average spectrum of clean speech matches produced speech in some parts, it is quite different in others. Therefore, since the goal is to obtain produced speech from device speech, intermediate decoloration of device speech to match clean speech could be unnecessary.

B. Denoising and Dynamic Range Compression

Dynamic range compression algorithms [21] are an essential part of the production process. They typically attenuate louder sounds in order to reduce the dynamic range of a recording and then amplify the entire signal in order to maintain the original level. This unfortunately amplifies background noise in addition to speech (Fig. 5). One can therefore consider a dynamic range compressor to invert the effect of a denoising algorithm to some degree. This is particularly noticeable in the parts of the recording between words. This can be circumvented to a degree by using a noise gate [2], [3] or voice activity detector [22], [23] and amplifying only parts with speech, but the noise floor will still be increased in some of these parts. It could therefore be beneficial to jointly consider denoising and dynamic range compression (rather than considering them as parts of independent subproblems) to attempt to reduce this issue.

C. Denoising and De-essing

Some fricatives of speech tend to be sibilant, which cause them to sound harsh. Effects such as dynamic range compression and equalization often exacerbate this harshness, which is undesirable [2]. Therefore, sound engineers often apply an effect called a de-esser, which attenuates sibilant

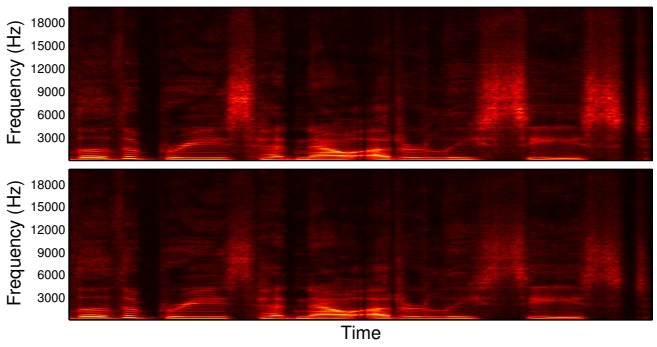


Fig. 6. Clean speech (top) has been processed by a de-esser (bottom). As shown the de-esser attenuates the fricatives.

sounds particularly in the 3-10 KHz range (Fig. 6). These sounds tend to be spectrally similar to wideband noise often found in device recordings. Therefore, denoising algorithms often attenuate sibilant sounds. This attenuation is typically considered undesirable when the goal is to recover clean speech. However, when the goal is to obtain produced speech, a greater degree of attenuation of sibilant sounds and therefore a more aggresive denoising technique could be acceptable.

IV. Evaluation Metrics

Several speech enhancement evaluation metrics exist in the literature [8], [10], [24], which gives us a way to evaluate estimated clean speech obtained from device speech. However, the right way to evaluate produced speech obtained from device speech or clean speech is less clear. Since there are aesthetic decisions involved in the creation of produced speech from clean speech, a number of solutions could be equally aesthetically pleasing and therefore equally correct. However, in order to make evaluation of the problem of obtaining produced speech from device speech more objective, we could simply determine how close the obtained produced speech is to the ground truth produced speech in this dataset. Since we are essentially trying to compute a form of a distance metric between two aligned clips of speech, we could potentially use certain existing speech enhancement metrics for this purpose.

Another approach could be to perform subjective listening tests and then develop objective metrics that are well correlated to these subjective results such as recently done in the case of audio source separation [25].

V. Conclusion

We have shown why it could be useful to transform device recordings into produced recordings, discussed insights and challenges with the problem, and described a new dataset that we have developed for the purpose of developing algorithms for this purpose. We believe that this dataset will help facilitate research into this problem, which is of growing importance.

Acknowledgements

We would like to thank Miik Dinko (the professional sound engineer who performed the recording and production) and the staff from Outpost Studios in San Francisco as well as all of the speakers who participated in the creation of the dataset.

REFERENCES

[1] B. Owsinski, *The Recording Engineer's Handbook*, 3rd ed. Cengage Learning, 2013.

[2] ——, *The Mixing Engineer's Handbook*, 3rd ed. Cengage Learning, 2013.

[3] A. Case, *Sound FX: Unlocking the Creative Potential of Recording Studio Effects*. Focal Press, 2007.

[4] B. Katz, *Mastering Audio: The Art and the Science*, 2nd ed. Focal Press, 2007.

[5] Y. Ephraim and D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 32, no. 6, December 1984.

[6] P. Scalart and V. Filho, "Speech enhancement based on a priori signal to noise estimation," in *Proceeding of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, May 1996.

[7] Z. Duan, G. J. Mysore, and P. Smaragdis, "Speech enhancement by online non-negative spectrogram decomposition in non-stationary noise environments," in *Proceedings of Interspeech*, September 2012.

[8] P. C. Loizou, *Speech Enhancement: Theory and Practice*, 2nd ed. CRC Press, 2013.

[9] P. Naylor and N. D. Gaubitch, *Speech Dereverberation*. Springer, 2010.

[10] K. Kinoshita, M. Delcroix, T. Yoshioka, T. Nakatani, E. Habets, R. Haeb-Umbach, V. Leutnant, A. Sehr, W. Kellermann, R. Maas, S. Gannot, and B. Raj, "The reverb challenge: A common evaluation framework for dereverberation and recognition of reverberant speech," in *Proceeding of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2013.

[11] D. Liang, D. P. Ellis, M. D. Hoffman, and G. J. Mysore, "Speech decoloration based on the product of filters model," in *Proceeding of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, May 2014.

[12] N. Enbom and B. Kleijn, "Bandwidth expansion of speech based on vector quantization of the mel frequency cepstral coefficients," in *Proceedings of the IEEE Workshop on Speech Coding*, June 1999.

[13] J. Han, G. J. Mysore, and B. Pardo, "Language informed bandwidth expansion," in *Proceedings of the IEEE International Workshop on Machine Learning for Signal Processing*, September 2012.

[14] V. Verfaille, U. Zölzer, and D. Arfib, "Adaptive digital audio effects (a-dafx): A new class of sound transformations," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 14, no. 5, September 2006.

[15] D. Giannoulis, M. Massberg, and J. D. Reiss, "Parameter automation in a dynamic range compressor," *Journal of the Audio Engineering Society*, vol. 61, no. 10, October 2013.

[16] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 8, 2013.

[17] C. E. Rasmussen and C. K. I. Williams, *Gaussian Processes for Machine Learning*. MIT Press, 2006.

[18] E. Vincent, J. Barker, S. Watanabe, J. L. Roux, F. Nesta, and M. Matassoni, "The second 'CHIME' speech separation and recognition challenge: Datasets, tasks, and baselines," in *Proceeding of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, May 2013.

[19] H.-G. Hirsch and D. Pearce, "The aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions," in *Proceedings of the ISCA workshop ASR2000*, September 2000.

[20] N. Parihar, J. Picone, D. Pearce, and H.-G. Hirsch, "Performance analysis of the aurora large vocabulary baseline system," in *Proceedings of the European Signal Processing Conference*, September 2004.

[21] D. Giannoulis, M. Massberg, and J. D. Reiss, "Digital dynamic range compressor design — a tutorial and analysis," *Journal of the Audio Engineering Society*, vol. 60, no. 6, June 2012.

[22] J. Sohn, N. S. Kim, and W. Sung, "A statistical model-based voice activity detection," *IEEE Signal Processing Letters*, vol. 6, no. 1, January 1999.

[23] F. G. Germain, D. Sun, and G. J. Mysore, "Speaker and noise independent voice activity detection," in *Proceedings of Interspeech*, August 2013.

[24] 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, January 2008.

[25] 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, 2011.