Experimental Study on Transfer Learning in Denoising Autoencoders for Speech Enhancement

Marvin Coto-Jiménez(✉)

University of Costa Rica, San José 11501-2060, Costa Rica
marvin.coto@ucr.ac.cr
https://eie.ucr.ac.cr/

Abstract. The quality of speech signals is affected by a combination of background noise, reverberation, and other distortions in real-life environments. The processing of such signals presents important challenges for tasks such as voice or speaker recognition. To enhance signals in such challenging conditions several deep learning-based methods have been proposed. Those new methods have proven to be effective, in comparison to classical algorithms based on statistical analysis and signal processing. In particular, recurrent neural networks, especially those with long short-term memory (LSTM and BLSTM), have presented surprising results in tasks related to enhancing speech. One of the most challenging aspects of artificial neural networks is to reduce the high computational cost of the training procedure. In this work, we present a comparative study on transfer learning to accelerate and improve traditional training based on random initialization of the internal weights of the networks. The results show the advantage of the proposal in terms of less training time and better results for the task of denoising speech signals at several signal-to-noise ratio levels of white noise.

Keywords: BLSTM · Deep learning · Transfer learning · Speech processing

1 Introduction

In real-life environments, audio signals are affected by various conditions, such as additive or convolutional noise due to elements that produce sounds simultaneously, obstacles in the signal path of the microphone or room size and materials. Communication devices and applications of speech technologies, such as speech recognition, may be affected in their performance [1–3] by the presence of any of these conditions.

To overcome this problem, many algorithms have been developed over the past few decades to enhance noisy speech. These algorithms try to reduce distortions, as well as improve the quality of the perceived signal. Several successful algorithms have been based on deep neural networks (DNN) [4].

Supported by University of Costa Rica. Project 322-B9-105.

K. M. Figueroa Mora et al. (Eds.): MCPR 2020, LNCS 12088, pp. 307–317, 2020.
https://doi.org/10.1007/978-3-030-49076-8_29

The benefits of achieving this type of speech signal enhancement can be applied to signal processing in mobile phone applications, robotics, voice over Internet protocol, speech recognition systems, and devices for people with diminished hearing ability [5,6].

Models of neural networks that have been successfully applied in denoising speech signals include recurrent neural networks (RNNs); e.g., the Long Short-term Memory (LSTM) neural networks and their bidirectional counterparts (BLSTM). In previous efforts to enhance speech, spectrum-derived characteristics, such as Mel-frequency cepstrum coefficients (MFCC), have been mapped successfully between noisy speech to clean speech [7–9].

The benefits of using LSTM and BLSTM arose from their superior modeling of the dependent nature of speech signals. In particular, LSTM has been applied for denoising speech signals, for applications such as speaker recognition [10]. A comparative study using several types of DNN has been presented in [11], concluding that the best benefits are obtained with LSTM, compared to classical DNN or convolutional networks. In spite of their advantages, one of the drawbacks often reported about LSTM and BLSTM is the high computational cost of their training procedures.

The concept of transfer learning for speech processing has been reported in applications such as speech synthesis [12], music genre classification [13] and robust speech recognition [14]. The results have shown improvement in signal quality, classification, and word error rate in these applications.

For the particular task of denoising speech, the transfer learning process has been tested with other kinds of networks; e.g., general adversarial networks [15]. In this work, the concept of transfer learning between neural networks is applied as a way to increase the effectiveness of BLSTM networks and reduce training time in the task of denoising signals at different signal-to-noise ratio (SNR) levels. To assess the improvements made, a comparative study on transfer learning from a single BLSTM network trained on a particular SNR level, as well as some other networks, is performed.

1.1 Problem Statement

A speech signal y degraded with additive noise from the environment, can be modeled as the sum of a clean speech signal, x, and noise d, given by:

$$y(t) = x(t) + d(t) \tag{1}$$

In classical methods, $x(t)$ is considered uncorrelated to $d(t)$. Many speech enhancement algorithms estimate the spectrum of $x(t)$ from the spectral domain of $y(t)$ and $d(t)$.

In deep learning approaches, $x(t)$ (or the spectrum $X_k(n)$) can be estimated using algorithms that learn an approximated function $f(\cdot)$ between the noisy and clean data of the form:

$$\hat{x}(t) = f\left(y(t)\right). \tag{2}$$

The precision of the approximation $f(\cdot)$ usually depends on the amount of training data and the algorithm selected.

In the approach presented in this work, the information for the training of the artificial neural network for a particular noise level is transferred from another network, so the approximation $f(\cdot)$ can be obtained with better precision and fewer training epochs, which means a significant increase in the efficiency.

For this purpose, several objective measures were used to verify the results, which comparatively show the benefits of the transfer learning for the denoising BLSTM neural networks. The rest of this document is organized as follows. Section 2 provides the background of BLSTM neural networks. Section 3 briefly describes the main definitions of Transfer Learning. Section 4 presents the experimental setup. In Sect. 5, the results and discussion are presented, and finally, Sect. 6 presents the conclusions.

2 BLSTM Neural Networks

Since the appearance of RNNs, new alternatives to model the character dependent on the sequential information have been presented. For example, the neural networks capable of storing information through feedback connections between neurons in their hidden layers or other neurons that are in the same layer [16,17].

One of the most important models are the LSTM networks shown in [18]. Here, a set of gates is introduced into internal units, making a system capable of controlling access, storage, and propagation of values across the network. The results obtained when using LSTM networks in areas that depend on previous states of information, as is the case with voice recognition, musical composition, and handwriting synthesis, were encouraging [19,20].

Specifically, each unit in the LSTM networks has additional gates for storing values: one for input, one for memory clearing, one for output, and one for activating memory. With the proper combination of the gates, it is possible to store values for many steps or have them available at any time [18].

The gates are implemented using the following equations:

$$i_t = \sigma\left(\mathbf{W}_{xi}x_t + \mathbf{W}_{hi}h_{t-1} + \mathbf{W}_{ci}c_{t-1} + b_i\right) \tag{3}$$

$$f_t = \sigma\left(\mathbf{W}_{xf}x_t + \mathbf{W}_{hf}h_{t-1} + \mathbf{W}_{cf}c_{t-1} + b_f\right) \tag{4}$$

$$c_t = f_t c_{t-1} + i_t \tanh\left(\mathbf{W}_{xc}x_t + \mathbf{W}_{hc}h_{t-1} + b_c\right) \tag{5}$$

$$o_t = \sigma\left(\mathbf{W}_{xo}x_t + \mathbf{W}_{ho}h_{t-1} + \mathbf{W}_{co}c_t + b_o\right) \tag{6}$$

$$h_t = o_t \tanh\left(c_t\right) \tag{7}$$

where σ is the sigmoid activation function, i is the input gate, f is the memory erase gate, o_t is the exit gate and h is the output of the LSTM memory unit. c is the activation of memory. $\mathbf{W}_{mn}$ is the matrix that contains the values of the connections between each unit and the gates. Additional details about the training process and the implications of this implementation can be found at [21].

An extension of LSTM networks that have had a greater advantage in tasks related to temporal parameter dependence is the Bidirectional LSTM (BLSTM). Here, the configuration of the network allows the updating of parameters in both directions during training.

LSTM networks can handle information over long periods; however, using BLSTM with two hidden layers connected to the same output layer gives them access to information in both directions. This allows bidirectional networks to take advantage of not just the past but also the future context [22].

One of the main architectures applied for speech enhancement is autoencoders. An autoencoder is a particular kind of neural network, which consists of an encoder that transforms an input vector s into a representation in the hidden layers h through a f mapping. It also has a decoder that takes the hidden representation and transforms it back into a vector in the input space. In implementations, the number of units at the inputs correspond to the number of units at the outputs of the network.

3 Transfer Learning

In the context of machine learning with artificial neural networks, transfer learning has been used to improve a model in one domain by transferring information from a model in a related domain. For example, given a source domain D_S with a corresponding task T_S, and a target domain D_T with a corresponding task T_T, transfer learning is the process of improving the target predictive function $f_T(\cdot)$ by using the related information from D_S and T_S [23].

Among the several types that have been developed, homogeneous transfer learning is the most suitable for the approach presented in this work. The homogeneous transfer is properly applied where there is available data that is drawn from a domain related to, but not an exact match for, a target domain of interest.

This process of transfer learning is commonly applied in human learning, where experiences in one task (e.g., playing a musical instrument) can mean a significant improvement in learning a new task (e.g., playing a different musical instrument), relative to learning a new task with no antecedent [24].

In this work, D_S and T_S represent several possibilities: the approximation of the identity function with speech data, the same task with noisy data, or the task of denoising speech with additive noise at SNR0, while D_T and T_T the task of denoising at every other SNR level.

4 Experimental Setup

This section describes in some detail the experimental setup that was followed in the paper. The whole process can be summarized in the following steps:

1. Noisy database generation: Files containing artificially generated white noise were generated and added to each audio file in the database for a given signal-to-noise ratio (SNR). Five noise levels were added, to cover a range from light to heavy noise levels.

2. Feature extraction and input-output correspondence: A set of parameters was extracted from the noisy and the clean audio files, using the Ahocoder system. Those from the noisy files were used as inputs to the networks, while the corresponding clean features were the outputs.
3. Training and validation: During training and validation, using forward pass and backpropagation through time algorithm to train the BLSTM networks, the internal weights of the connections were adjusted as the noisy and clean utterances were presented at the inputs and the outputs. A total of 900 utterances (about 80% of the total database) were used for training and 180 utterances (about 15% of the total database) were used for validation. Details and equations of the algorithm followed can be found in [25].
4. Test: A subset of 50 randomly selected utterances (about 5% of the total amount of utterances of the database) was chosen for the test set. These utterances were not part of the training process, to provide independence between the training and testing.

4.1 Database

In our work, we chose the SLT voice from the CMU ARCTIC databases [26], designed for speech research purposes at the Carnegie Mellon University, in the United States of America. The whole set of 1132 sentences were used to randomly define the training, validation and test sets. In our work, we chose the female SLT voice.

4.2 Initialization and Transfer Learning

To test the proposal, different proposals to initialize the networks, or perform the transfer learning from other networks are compared:

- Base system (random initialization): The set of internal weights of the BLSTM network were initialized randomly, as the common practice in training autoencoders.
- Transfer (AA-clean): In this approach, an auto-associative network (which approximate the identity function between the inputs and the outputs) is trained using the MFCC from the clean speech. The result of this previous training procedure is applied as the initialization weights of the denoising autoencoders for each SRN.
- Transfer (AA-noisy): Similar to the previous case, an auto-associative network is previously trained, but using parameters from the noisy speech for each SNR level. The result of this training procedure is applied as the initialization weights of the denoising autoencoders for each SRN.
- Transfer (TN): A denoising autoencoder was first randomly initialized, and then trained to denoise the parameters of the speech at SNR 0. The resulting weights of this network are then used as the initialization weights of the networks for every other SNR level.

The comparative study aims to find the best option to reduce training time and achieve better results in the denoising task, using the evaluation measures presented in the following section.

4.3 Evaluation

To assess the different to evaluate the results given by the different enhancement methods:

- PESQ: This measure uses a model to predict the perceived quality of speech. As defined in the ITU-T recommendation P.862.ITU, the results are given in interval $[0.5, 4.5]$, where 4.5 corresponds to a perfect reconstruction of the signal [27].
- SSE (Sum of squared errors): This is a common measure to evaluate the lowest validation error during training. SSE is defined as:

$$\mathrm{SSE}(\theta) = \sum_{n=1}^{T} (\mathbf{c_x} - \hat{\mathbf{c}}_\mathbf{x})^2 \tag{8}$$

  where $\mathbf{c_x}$ is the desired output of the network, $\hat{\mathbf{c}}_\mathbf{x}$ is the obtained output, and T the number of frames.
- Number of epochs: During training the BLSTM networks, each epoch consists of a feedforward and backforward step to adjust the weights of the internal connections. The time taken to train the BLSTM is directly associated with the number of epochs in training.

5 Results and Discussion

In Table 1, the results of the training process for each of the initializations and transfer learning are presented. As stated in Sect. 3, the information from the BLSTM autoencoder trained with SNR 0 was transferred to the rest of the autoencoders. This is the reason for the missing values in the corresponding line of the SNR 0 measures.

Due to the random initialization of the networks in the base system, these training procedures were performed three times. Thus, the values reported in Table 1 correspond to the mean values for the base system.

Regarding the efficiency of the training procedure, transfer learning represents a significant advantage over most of the other approaches compared in this study. For example, for SNR -10 (the heavier level of white noise), the training time is reduced by more than 55% in comparison to the base system.

For SNR 10, the training time is reduced 20%, while the reductions for SNR 5 and SNR -5 are 50% and 52% respectively. This represents a significant improvement in efficiency in all cases. The benefits of using transfer learning from a particular SNR level is also clear when compared with the other approaches for the initialization of the BLSTM network. Considering that each epoch requires

Table 1. Average number of epochs for training, SSE and PESQ for training the BLSTM denoise autoencoder with white noise at several SNR. * is the best result for each type of noise and each measure.

SNR 10			
Training model	Avg. epochs	SSE	PESQ
Base system	526	244.49	2.58
Transfer (TN)	417	245.95	2.57
Transfer (AA-clean)	655	243.79*	2.59*
Transfer (AA-noisy)	381*	248.14	2.57
SNR 5			
Training model	Avg. epochs	SSE	PESQ
Base system	318	288.36	2.26
Transfer (TN)	157*	289.39	2.28*
Transfer (AA-clean)	394	283.22	2.27
Transfer (AA-noisy)	508	279.09*	2.27
SNR 0			
Training model	Avg. epochs	SSE	PESQ
Base system	165*	335.13	1.75
Transfer (TN)	–	–	–
Transfer (AA-clean)	374	328.28	1.77*
Transfer (AA-noisy)	328	325.7*	1.76
SNR −5			
Training model	Avg. epochs	SSE	PESQ
Base system	284	387.58	0.96*
Transfer (TN)	136*	342.3*	0.96*
Transfer (AA-clean)	270	392.07	0.96*
Transfer (AA-noisy)	246	387.12	0.96*
SNR −10			
Training model	Avg. epochs	SSE	PESQ
Base system	224	462.04	0.47
Transfer (TN)	96*	447.94*	0.55
Transfer (AA-clean)	196	463.06	0.56*
Transfer (AA-noisy)	206	457.22	0.52

about 60 s in a desktop computer accelerated with an NVIDIA GPU, several hours can be saved for the whole set of experiments.

In terms of SSE, the reduction in the number of epochs is also reflected in better values in two of the four cases compared. For SNR 10 and SNR 5 (the lightest levels of noise), the best SSE value was obtained with the

auto-associative initialization with clean and noisy MFCC values. The values obtained with the training performed after the transfer learning do not differ significantly from those of the best case.

Finally, in terms of the PESQ, transfer learning reaches the best case in two of the four cases applied. At all levels, except for SNR 10, there is an increase in the PESQ measure in comparison to the base system.

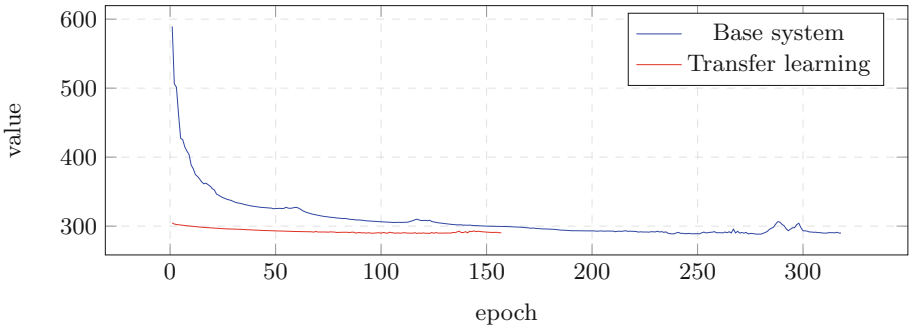

Fig. 1. Evolution of the SSE during training, for SNR5

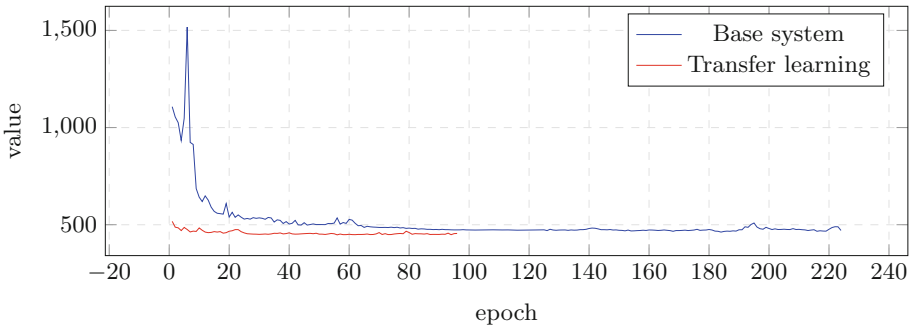

Fig. 2. Evolution of the SSE during training, for SNR-10

With all this information, the benefits of transfer learning in denoising applications, using BLSTM autoencoders, are clear. The training time drops significantly and the objective measures of quality also increase, or do not differ significantly from the best case.

In Fig. 1 and 2 the evolution of the validation error during training is presented. It is remarkable how much less time it takes for the training procedure to reach the lower SEE value. With this reduction in training time and the similar or best results in the other measures, a greater number of experiments or cases can be analyzed using this type of artificial neural network.

6 Conclusions

In this work, a comparative study of four approaches to initialize BLSTM autoencoders was presented. In particular, the main focus was on the transfer learning approach, which brings a set of weights adjusted after a training procedure for a particular level of white noise to other levels of noise.

Transfer learning presents benefits in terms of efficiency during training, in comparison to two other supervised initializations (in the form of auto-associative memories) and the more traditional random initialization approach.

The reduction of training time for this kind of network, with a large number of connections, can be measured in terms of hours or even days in a large set of experiments, such as the one performed in this study.

For future work, more extensive research about the source of the transfer learning (such as the SNR used) can be performed. Statistical validation of the improvements achieved could be also relevant. Finally, additional benefits about transfer learning for different kinds of noise can be of great interest to speech enhancement applications.

References

1. Weninger, F., et al.: Deep recurrent de-noising auto-encoder and blind de-reverberation for reverberated speech recognition. In: 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE (2014)
2. Donahue, C., Bo, L., Prabhavalkar, R.: Exploring speech enhancement with generative adversarial networks for robust speech recognition. In: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE (2018)
3. Coto-Jiménez, M., Goddard-Close, J., Martínez-Licona, F.: Improving automatic speech recognition containing additive noise using deep denoising autoencoders of LSTM networks. In: Ronzhin, A., Potapova, R., Németh, G. (eds.) SPECOM 2016. LNCS (LNAI), vol. 9811, pp. 354–361. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-43958-7_42
4. Abouzid, H., Chakkor, O., Reyes, O.G., Ventura, S.: Signal speech reconstruction and noise removal using convolutional denoising audioencoders with neural deep learning. Analog Integr. Circ. Sig. Process **100**(3), 501–512 (2019). https://doi.org/10.1007/s10470-019-01446-6
5. Lai, Y.-H., et al.: A deep denoising autoencoder approach to improving the intelligibility of vocoded speech in Cochlear implant simulation. IEEE Trans. Biomed. Eng. **64**(7), 1568–1578 (2016)
6. Coto-Jimenez, M., Goddard-Close, J., Di Persia, L., Rufiner, H.L.: Hybrid speech enhancement with wiener filters and deep LSTM denoising autoencoders. In: Proceedings of the 2018 IEEE International Work Conference on Bioinspired Intelligence (IWOBI), San Carlos, CA, USA, 18–20 July 2018, pp. 1–8 (2018)
7. Gutiérrez-Muñoz, M., González-Salazar, A., Coto-Jiménez, M.: Evaluation of mixed deep neural networks for reverberant speech enhancement. Biomimetics **5**(1), 1 (2020)

8. Chakraborty, R., et al.: Front-end feature compensation and denoising for noise robust speech emotion recognition. In: Proceedings of Interspeech 2019, pp. 3257–3261 (2019)
9. Coto-Jiménez, M.: Robustness of LSTM neural networks for the enhancement of spectral parameters in noisy speech signals. In: Batyrshin, I., Martínez-Villaseñor, M.L., Ponce Espinosa, H.E. (eds.) MICAI 2018. LNCS (LNAI), vol. 11289, pp. 227–238. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-04497-8_19
10. Tkachenko, M., Yamshinin, A., Lyubimov, N., Kotov, M., Nastasenko, M.: Speech enhancement for speaker recognition using deep recurrent neural networks. In: Karpov, A., Potapova, R., Mporas, I. (eds.) SPECOM 2017. LNCS (LNAI), vol. 10458, pp. 690–699. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-66429-3_69
11. Liu, M., et al.: Speech enhancement method based on LSTM neural network for speech recognition. In: 2018 14th IEEE International Conference on Signal Processing (ICSP). IEEE (2018)
12. Jia, Y., et al.: Transfer learning from speaker verification to multi speaker text-to-speech synthesis. In: Advances in Neural Information Processing Systems (2018)
13. Song, G., Wang, Z., Han, F., Ding, S.: Transfer learning for music genre classification. In: Shi, Z., Goertzel, B., Feng, J. (eds.) ICIS 2017. IAICT, vol. 510, pp. 183–190. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-68121-4_19
14. Jiangyan, Y.I., et al.: Transfer learning for acoustic modeling of noise robust speech recognition. J. Tsinghua Univ. (Sci. Technol.) **58**(1), 55–60 (2018)
15. Pascual, S., et al.: Language and noise transfer in speech enhancement generative adversarial network. In: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE (2018)
16. Fan, Y., Qian, Y., Xie, F.L., Soong, F.K.: TTS synthesis with bidirectional LSTM based recurrent neural networks. In: Proceedings of the Fifteenth Annual Conference of the International Speech Communication Association, Singapore, 14–18 September (2014)
17. Zen, H., Sak, H.: Unidirectional long short-term memory recurrent neural network with recurrent output layer for low-latency speech synthesis. In: Proceedings of the 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), South Brisbane, Australia, 19–24 April 2015, pp. 4470–4474 (2015)
18. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Comput. **9**, 1735–1780 (1997)
19. Graves, A., Jaitly, N., Mohamed, A.R.: Hybrid speech recognition with deep bidirectional LSTM. In: Proceedings of the 2013 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU), Olomouc, Czech Republic, 8–13 December 2013, pp. 273–278 (2013)
20. Graves, A., Fernández, S., Schmidhuber, J.: Bidirectional LSTM networks for improved phoneme classification and recognition. In: Duch, W., Kacprzyk, J., Oja, E., Zadrożny, S. (eds.) ICANN 2005. LNCS, vol. 3697, pp. 799–804. Springer, Heidelberg (2005). https://doi.org/10.1007/11550907_126
21. Gers, F.A., Schraudolph, N.N., Schmidhuber, J.: Learning precise timing with LSTM recurrent networks. J. Mach. Learn. Res. **3**, 115–143 (2002)
22. Wöllmer, M., Eyben, F., Schuler, B., Rigoll, G.: A multi-stream ASR framework for BLSTM modeling of conversational speech. In: Proceedings of the 2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Prague, Czech Republic, 22–27 May 2011, p. 4861 (2011)
23. Weiss, K., Khoshgoftaar, T.M., Wang, D.D.: A survey of transfer learning. J. Big Data **3**(1), 1–40 (2016). https://doi.org/10.1186/s40537-016-0043-6

24. Pan, S.J., Yang, Q.: A survey on transfer learning. IEEE Trans. Knowl. Data Eng. **22**(10), 1345–1359 (2009)
25. Greff, K., et al.: LSTM: a search space odyssey. IEEE Trans. Neural Netw. Learn. Syst. **28**(10), 2222–2232 (2016)
26. Kominek, J., Black, A.W.: The CMU Arctic speech databases. In: Fifth ISCA Workshop on Speech Synthesis (2004)
27. Rix, A.W., et al.: Perceptual evaluation of speech quality (PESQ) the new ITU standard for end-to-end speech quality assessment Part I-time-delay compensation. J. Audio Eng. Soc. **50**(10), 755–764 (2002)