Automatic Speech Recognition Performance for Training on Noised Speech

Arkadiy Prodeus
Acoustics and Electroacoustics Department, National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute"
Kyiv, Ukraine
aprodeus@gmail.com

Kateryna Kukharicheva
Acoustics and Electroacoustics Department, National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute"
Kyiv, Ukraine
katerynakt@gmail.com

***Abstract*— Performances of some training techniques of automatic speech recognition system are compared in this paper. Speech recognition accuracy was used as measure of performance. Different kinds of outdoor and indoor noise were used for studying. It is shown the superiority of training on noised speech methods over the competitive technique of training on clear speech. It has been found that training by means of noised speech allows reach high, abut 95…97%, recognition accuracy for about 5…10 dB signal-to-noise ratio. At the same time, training by means of clean speech allows reach the same accuracy for about 20…25 dB.**

Keywords—noised speech; clean speech; automatic speech recognition performance; training technique

I. Introduction

Nowadays, automatic speech recognition (ASR) systems robust to action of noise and reverberation are much claimed. That is why speech enhancement system as ASR pre-processors for noise and late reverberation reduction (Fig. 1) are often used in different applications, and communication, PC and smartphone applications [1-4] are among them.

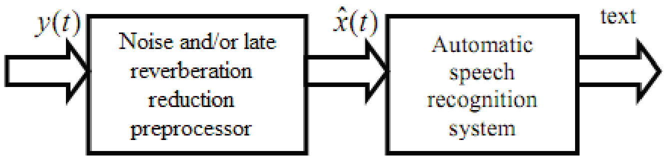

Fig. 1. Speech enhancement system as ASR pre-processor

Another way to improve the ASR systems robustness is to change the way of training of the systems. One can points different approaches to train ASR systems for operation in a noisy rooms and streets [3]. Usually ASR systems are trained using clean speech. But ASR systems can also be trained using noisy speech. As it was shown, the second approach provides essentially higher recognition accuracy [3–7], i.e. a much higher ASR system robustness.

Under second approach, four training techniques are possible and promising for applications. They are shown in Table 1, where SNR_t is signal-to-noise ratio (SNR) in the training mode and SNR_r is SNR in the recognition mode, $n_t(t)$ and $n_r(t)$ are background noises in proper modes, respectively.

TABLE I. Noised Speech Training Techniques

Technique	Matching
Fully matched training (FMT)	$\mathrm{SNR}_t = \mathrm{SNR}_r$, $n_t(t) = n_r(t)$
Noise matched training (NMT)	$\mathrm{SNR}_t \neq \mathrm{SNR}_r$, $n_t(t) = n_r(t)$
SNR matched training (SNRMT)	$\mathrm{SNR}_t = \mathrm{SNR}_r$, $n_t(t) \neq n_r(t)$
Multi-style training (MST)	$\mathrm{SNR}_t \neq \mathrm{SNR}_r$, $n_t(t) \neq n_r(t)$

When ASR systems are trained and tested on speech with equal SNR and the same kinds of noise, we can say about "fully matched training" (FMT) technique. Evident disadvantage of the FMT technique is strong requirement of huge ASR system storage for remembering of phonemes definitions for all noise and SNR combinations. At the same time, this technique is high-performance: speech recognition accuracy ($Acc\%$) is close to 75% for SNR = 5 dB whereas $Acc\%$ is about 25% for training on clean speech [3]. Because of these results belong to a particular case of training on white noise, elimination of this drawback was realized in [4] where 14 kinds of noises were studied (Fig. 2). These results proper to case of street paved with stone blocks, but similar results were obtained for all 14 kinds of considered noises. Moreover, these results are well matched with ones of [3] and are interesting because characterize the FMT technique more fully.

Indeed, for clean speech training technique (Fig. 3), recognition accuracy $Acc\% \approx 95\%$ was achieved for $\mathrm{SNR}_r > 28$ dB in noise case of paved street. Meanwhile, recognition accuracy $Acc\% \approx 95\%$ was achieved upon $\mathrm{SNR}_r = 7 \ldots 15$ dB and $\mathrm{SNR}_t \approx 10$ dB for FMT technique. Value $Acc\% \approx 95\%$ can be achieved for $\mathrm{SNR}_r \approx 8 \ldots 27$ dB

978-1-5386-0637-7/17/$31.00 ©2017 IEEE

and $\text{SNR}_t \approx 15$ dB. Rise of SNR_t to 20 dB demands $\text{SNR}_r \approx$ 12...35 dB to reach $Acc\% \approx 95\%$. Evidently, SNR_t rise leads to extension and moves right SNR_r values range that guarantee high recognition accuracy.

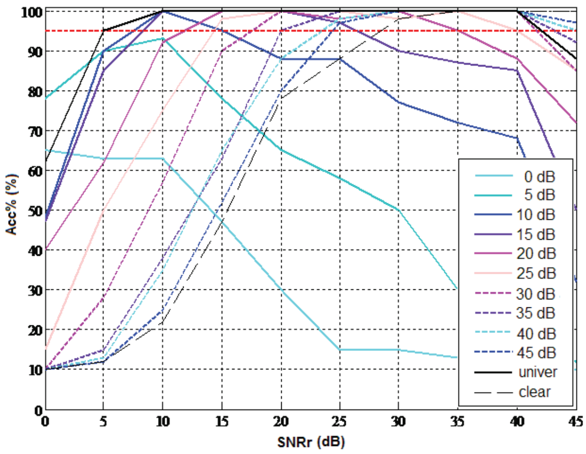

Fig. 2. Estimates of Acc% for FMT technique (paved street) [4]

Thus, second evident disadvantage of the FMT technique is phenomenon of $Acc\%$ decreasing upon SNR_r rise since dependence $Acc\%(SNR_r)$ has marked maximum for $\text{SNR}_r \approx \text{SNR}_t$.

Technique for case $n_t(t) = n_r(t)$ but $\text{SNR}_t \neq \text{SNR}_r$ considered in [8] can be named as "noise matched training" (NMT). Evidently, it can be useful for situations when noise kind is a priori known. In contrast to FMT technique, NMT technique is much less strict to ASR system storage. Unfortunately, there is no quantitative assessment of its performance in the literature. Thus, filling of this gap is one of the objectives of this paper.

Case of $n_t(t) \neq n_r(t)$ but $\text{SNR}_t = \text{SNR}_r$ can be named "SNR matched training" (SNRMT) technique. This technique is reasonable when a priori information about noise kind is absent at recognition mode. Thus, the technique is much more interesting for practice compare to NMT technique because ASR system can be easily moved from one place to another if it will be needed. Disadvantage of the technique is high training time because a lot of different noise kinds need be used.

But the most interesting and very promising is "multi-style training" (MST) technique [4] corresponding to situation of training with all available noises and signal-to-noise ratios. As can be seen, this technique is intelligent in an unknown environment. Of course, MST technique has some disadvantages: 1) extremely high training time; 2) inability to use in training mode all noise and SNR combinations that can occur in recognition mode. Of course, the same disadvantages are peculiar to SNRMT technique too. But essential advantage is much lighter ASR system storage requirement for MST technique what has great practical value. It was found that $Acc\%$ is almost equal for FMT and MST techniques [3, 6]. It has been claimed in [7] that MST technique exceeds clean speech training technique (about 20%) when speech enhancing pre-processor is used in recognition mode. But a set of noises considered in [7] was limited (office and car). Thus, another object of this paper is to eliminate this shortcoming.

II. Experiment Organization

In this paper, FMT, NMT and MST techniques have been compared among themselves and with clean speech training technique.

Additive mixtures of signal and noise were formed:

$$s(t) = k \cdot x(t) + n(t), \quad k = 10^{0.05(SNR_0 - SNR)},$$

$x(t)$ is clear speech signal, $n(t)$ is noise signal, SNR_0 is desired signal-to-noise ratio, SNR corresponds to initial speech and noise signals. SNR_0 values were varied between 0 and 45 dB.

Names of ten numbers (in Russian) were used as speech signals. Used fourteen kinds of noises (Table II) can by grouped in three sets. First, it is group of indoor noises produced by home and office equipment: washer, grinder, microwave, and computer. Second group is street and transport noises: paved street, truck, subway train. Third group contain indoor and outdoor noises and thus have some signs of first two groups but also contain people conversation noise. They are filled audience and underpass noises, station and metro lobbies, places near station and trolleybus stop, in trolley noises.

HTK toolkit was used for simulation of ASR system and assessment of its recognition accuracy [7]. Twenty two Russian language phonemes were used in phoneme vocabulary. Thirty nine MFCC_0_D_A coefficients has been used upon ASR simulation. Single words of clean speech (SNR was near 45 dB) were recorded in anechoic room with 0.1 s reverberation time. Sampling rate of saved speech was 22050 Hz and 16 bit linear quantization was used. Every recorded word was uttered 20 times with a different intonation by single speaker-woman.

Test sentences contained ten words paused by 0.3–0.5 s. Six samples of noisy sentences were used for recognition accuracy assessment in accordance with equation

$$Acc\% = (N - D - S - I)/N \times 100\%,$$

D, S and I are numer of deletion, substitution and insertion errors, respectively; N is the total number of labels of the reference phrase.

III. Experimental Results

$Acc\%$ estimates for clean speech training technique are shown in Fig. 3 and Table II. As can be seen, recognition accuracy essentially depends on the temporal and spectral noise properties.

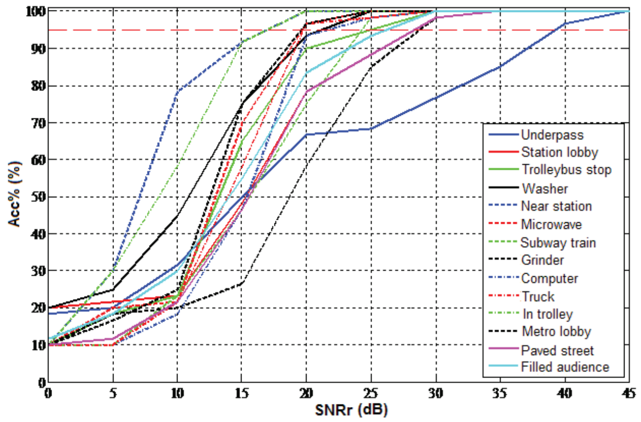

Fig. 3. Estimates of Acc% for clean speech training technique

TABLE II. ESTIMATES OF ACC% FOR CLEAN SPEECH TRAINING

Noises	SNR_r (dB)								
	0	**5**	**10**	**15**	**20**	**25**	**30**	**35**	**40**
Grinder	10	18	20	27	58	85	98	100	100
Computer	10	10	18	47	93	98	100	100	100
Microwave	10	20	22	70	97	100	100	100	100
Washer	20	25	45	75	93	100	100	100	100
Paved street	10	12	22	47	78	88	98	100	100
Subway train	10	10	23	47	75	98	100	100	100
Truck	10	10	22	58	97	98	100	100	100
In trolley	10	30	58	92	100	100	100	100	100
Filled audience	12	18	30	55	83	93	100	100	100
Trolleybus stop	10	18	23	65	90	95	100	100	100
Near station	10	30	78	92	100	100	100	100	100
Station lobby	20	22	23	48	78	88	98	100	100
Metro lobby	10	17	25	75	97	100	100	100	100
Underpass	18	20	32	50	67	68	77	85	97

For example, $Acc\% \approx 95\%$ for $SNR_r > 17$ dB speech in trolley, and $Acc\% \approx 95\%$ for $SNR_r > 25$ dB speech in filled audience. Underpass noise is the most dangerous when high recognition accuracy is demanded. This phenomenon can be explained as result of combined masking action of noise and reverberation interferences [10–12].

Results of testing of ASR system trained in accordance with NMT technique are shown in Fig. 4 and Table III. Comparison of NMT and FMT techniques allows one to give preference to the NMT method. For example, value $Acc\% = 95\%$ was reached for $SNR_r \geq 5$ dB for case of paved street. In most other cases the same accuracy was reached for $SNR_r > 10$ dB.

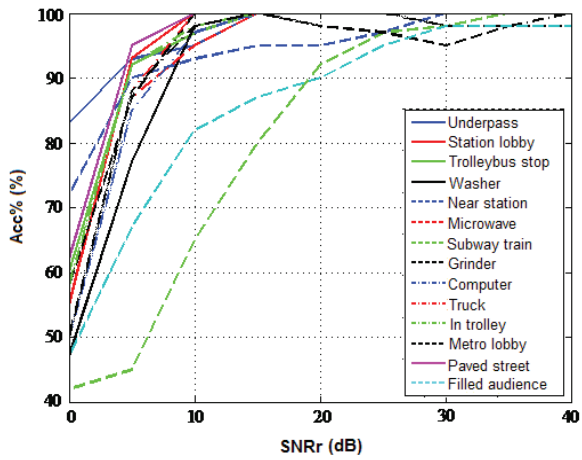

Fig. 4. Estimates of Acc% for NMT technique

TABLE III. ESTIMATES OF ACC% FOR NMT TECHNIQUE

Noises	SNR_r (dB)								
	0	**5**	**10**	**15**	**20**	**25**	**30**	**35**	**40**
Grinder	58	87	100	100	98	97	95	98	98
Computer	50	85	97	100	100	100	100	100	100
Microwave	58	87	95	100	100	100	100	100	100
Washer	47	77	98	100	100	100	100	100	100
Paved street	62	95	100	100	100	100	100	100	100
Subway train	42	45	65	80	92	97	98	100	100
Truck	50	88	100	100	100	100	100	100	100
In trolley	58	92	98	100	100	100	100	100	100
Filled audience	47	67	82	87	90	95	98	98	98
Trolleybus stop	60	92	97	100	100	100	100	100	100
Near station	72	90	93	95	95	97	100	100	100
Station lobby	55	93	100	100	100	100	100	100	100
Metro lobby	50	88	98	100	100	100	98	98	100
Underpass	83	93	95	100	100	100	100	100	100

Only noise in the people filled auditorium and subway train noise were exceptions for which $Acc\% = 95\%$ was reached only for $SNR_r > 25$ dB. Great advantage is also absence of abovementioned maximum of dependence $Acc\%(SNR_r)$ for $SNR_r \approx SNR_t$.

Results for MST technique are shown in Fig. 5 and Table IV. As can be seen, recognition accuracy $Acc\% \geq 97\%$ for $SNR_r \geq 10$ for most types of noise, which is even slightly better than for NMT technique. For $0 \leq SNR_r < 10$ dB, recognition accuracy is higher than ones for education on "clean" signals, but worse than ones for FMT and NMT techniques. There is special case of grinder noise which is much worse than other considered noises. Generally, given the much smaller requirements of MST technique to ASR system storage, we can consider the technique as preferable for $SNR_r \geq 10$ dB dB.

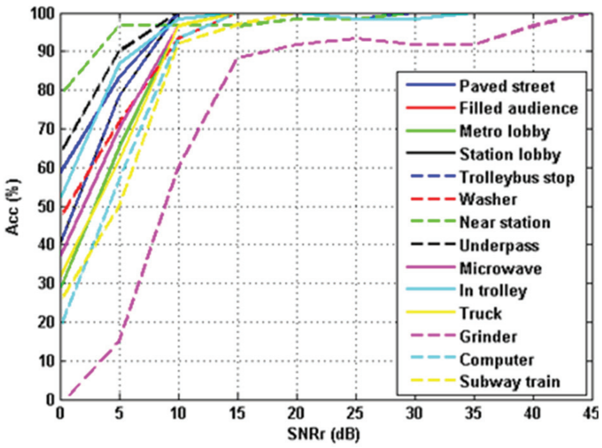

Fig. 5. Estimates of Acc% for MST technique

TABLE IV. ESTIMATES OF ACC% FOR MST TECHNIQUE

Noises	SNR_r (dB)								
	0	5	10	15	20	25	30	35	40
Grinder	-2	15	60	88	92	93	92	92	97
Computer	18	57	93	100	100	100	100	100	100
Microwave	37	70	97	100	100	100	100	100	100
Washer	47	72	93	100	100	100	100	100	100
Paved street	58	83	100	100	100	98	100	100	100
Subway train	25	50	92	97	100	100	100	100	100
Truck	32	62	97	100	100	100	100	100	100
In trolley	52	87	98	100	100	98	98	100	100
Filled audience	28	65	97	100	100	100	100	100	100
Trolleybus stop	40	78	100	100	100	100	100	100	100
Near station	78	97	97	97	98	98	100	100	100
Station lobby	40	78	100	100	100	100	100	100	100
Metro lobby	28	65	97	100	100	100	100	100	100
Underpass	83	93	95	100	100	100	100	100	100

CONCLUSION

Performances of some training techniques of automatic speech recognition system are compared among themselves and with clean speech training technique.

It is shown that NMT and MST techniques allow reach high, about 95…97%, recognition accuracy for $SNR_r > 5\ldots10$ dB. At the same time, training by means of clean speech allows reach the same accuracy for $SNR_r \geq 20$ dB. FMT technique has high performance too, but essential disadvantages of the technique are strong requirement to ASR storage and phenomenon of recognition accuracy decreasing when signal-to-noise is increasing.

Thus, ASR performance for noised speech training techniques was estimated and its superiority over clean speech training technique was experimentally proved.

REFERENCES

[1] V. Didkovskyi, S. Naida. “Building-up principles of auditory echoscope for diagnostics of human middle ear,” Radioelectronics and Communications Systems, 2016, V. 59, No. 1, p. 39-46. DOI: 10.3103/S0735272716010039

[2] S. Naida, “Acoustic theory problems of speech production in the light of the discovery of the formula for the middle ear norm parameter,” Proceedings of IEEE 35th International Conference on Electronics and Nanotechnology (ELNANO), pp. 347-350, 21-24 April 2015, Kyiv, Ukraine. DOI: 10.1109/ELNANO.2015.7146907 .

[3] X. Huang, A. Acero, and H.-W.Hon, Spoken Language Processing: a Guide to Theory, Algorithm, and system development. Prentice Hall, Inc., 2001, 965 p.

[4] A. Prodeus, K. Kukharicheva, “Training of automatic speech recognition system on noised speech,” Proc. 2016 IEEE 4th Int. Conf. "Methods and Systems of Navigation and Motion Control (MSNMC)," October 18-20, 2016, Kyiv, Ukraine. - P. 221-223.

[5] A. Prodeus, K. Kukharicheva, "Accuracy of Automatic Speech Recognition System Trained on Noised Speech," Electronics and Control Systems, No.3(49), 2016. – P. 11-16. ISSN: 1990-5548.

[6] R.P. Lippmann, E.A. Martin, and D.P. Paul, "Multi-Style Training for Robust Isolated-Word Speech Recognition," Int. Conf. on Acoustics, Speech and Signal Processing, pp. 709-712, 1987, Dallas, TX.

[7] J. Rajnoha, “Multi-Condition Training for Unknown Environment Adaptation in Robust ASR Under Real Conditions,” Acta Polytechnica vol. 49, no. 2–3, pp. 3-7, 2009.

[8] J. Li, L. Deng, Y. Gong, and R. Haeb-Umbach, “An Overview of Noise-Robust Automatic Speech Recognition,” IEEE/ACM Trans. Audio, Speech, and Language Processing, vol. 22, no. 4, pp. 745-777, February 2014.

[9] The HTK Book / Ed. S. Young, G. Evermann, M. Gales. Cambridge: University Engineering Department, 2009, 375 p.

[10] A. Prodeus and V.P. Ovsianyk, “Estimation of late reverberation spectrum: Optimization of parameters,” Radioelectronics and Communications Systems, vol. 58, Is. 7, pp.322-328, July 2015.

[11] V.S. Didkovskyi, S.A. Naida, and O.A. Zubchenko, “Technique for rigidity determination of the materials for ossicles prostheses of human middle ear,” Radioelectronics and Communications Systems, vol. 58, no. 3, pp. 134-138, 2015.

[12] K. Pylypenko, A. Prodeus, “Noise Impact Assessment on the Accuracy of the Determination of Speaker’s Gender by Using Method of the Cumulant Coefficients,” XIth International Conference "Perspective Technologies and Methods in MEMS Design (MEMSTECH 2015), Lviv–Polyana, Ukraine, pp. 102-106, 2-6 September 2015.