Training of Automatic Speech Recognition System on Noised Speech

Arkadiy Prodeus
Acoustics and Electroacoustics Department
Faculty of Electronics, NTUU KPI
Kyiv, Ukraine
aprodeus@gmail.com

Kateryna Kukharicheva
Acoustics and Electroacoustics Department
Faculty of Electronics, NTUU KPI
Kyiv, Ukraine
katerynakt@gmail.com

Abstract—In this paper, two techniques of automatic speech recognition system training on noised speech are compared with technique of training on clean speech. The comparing has been made by means of speech recognition accuracy measure, with usage of fourteen kinds of noise. These were noises of household appliances and computers, street and transport, teaching rooms and lobbies. The superiority degree of noised speech training techniques over the competitive technique has been assessed. It is shown that training on noised speech allows reaching the 95% recognition accuracy for minimal signal-to-noise ratio 10 dB, whereas training on clean speech allows reaching the same recognition accuracy for minimal signal-to-noise ratio 20 dB.

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

I. INTRODUCTION

A number of new aviation systems are beginning to utilize elements of artificial intellect. The F-35 was the first U.S. fighter aircraft with automatic speech recognition (ASR) system able to "hear" a pilot's spoken commands to manage various aircraft subsystems, such as communications and navigation [1]. It is believed also that voice control would enable air battle managers to control their unmanned aerial vehicles (UAVs) using voice commands in addition to joystick, mouse, and keyboard inputs [2]. But ambient cockpit noise or battle noise degrades the quality of the spoken command entering the recognition system, which could cause the system to misinterpret or misunderstand the command. Therefore developing of noise-robust ASR systems is present-day issue.

It can be pointed two approaches to training of ASR systems operating in a noisy environment [3]. In the first approach, the ASR system is trained on clean speech, when in the second approach the ASR system is trained on noisy speech. Studies show that the second approach is able to provide a much higher recognition accuracy [3] – [6].

Under this approach, three training techniques are most interesting for engineering applications. They are presented in Table 1, where SNR_t and SNR_r are signal-to-noise ratio in the training and recognition, respectively, $P_{nt}(f)$ and $P_{nr}(f)$ are noise spectrums in the training and recognition, respectively.

When "fully matched training" (FMT) method is used, ASR system is trained on speech with the same SNR and noise spectrum for which ASR system will be tested. As it is shown in [3], FMT method is very effective: for SNR = 5 dB, speech recognition accuracy $Acc\%$ =75%, whereas $Acc\%$ =25% for clean speech training [3]. Unfortunately, results given in [3] are limited to a special case of the discrete white noise. Therefore one of the objectives of this work is to eliminate this drawback.

In accordance with "multi-style training" (MT) technique [4], training is realized with all available noisy speech data. MT and FMT techniques are almost equal in terms of the $Acc\%$ [3], [4], and MT technique is about 20% better than the ASR system with training on clean speech and noise suppression at recognition [5]. As stated in [5], the recognition accuracy can be increased by 30% if the MT technique would be supplemented with noise suppression at recognition. Another important advantage of the MT technique is that it is much less demanding on memory size of ASR system.

A significant disadvantage of MT technique is the inability to use, when training, all combinations of noise kinds and SNR values that may occur during recognition. Therefore, it was proposed in [6] to produce training with varying SNR for noise that will affect the ASR system during recognition. This technique can be called "spectrum matched training" (SMT) (Table I). Although the rationality of this method is beyond doubt, there are no quantitative assessments of its effectiveness in the literature. Therefore, another objective of this paper is to fill this gap.

TABLE I. TRAINING TECHNIQUES

Technique name	Matching
Fully matched training (FMT)	$\mathrm{SNR}_t = \mathrm{SNR}_r$, $P_{nt}(f) = P_{nr}(f)$
Multi-style training (MT)	$\mathrm{SNR}_t \neq \mathrm{SNR}_r$, $P_{nt}(f) \neq P_{nr}(f)$
Spectrum matched training (SMT)	$\mathrm{SNR}_t \neq \mathrm{SNR}_r$, $P_{nt}(f) = P_{nr}(f)$

978-1-5090-1053-0/16/$31.00 ©

II. Problem Statement and Experiment Organization

ASR systems trained on clean and noisy signals were compared, on recognition accuracy, in this study. In this connection two techniques, FMT and SMT, of ASR system training on noisy signals were considered.

Additive mixture of signal and noise with desired signal-to-noise ratio SNR_0 was formed in accordance with equation:

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

where $x(t)$ is clear speech signal, $n(t)$ is noise, SNR is signal-to-noise ratio for saved clear speech signal. SNR_0 value was varied in the range 0–45 dB.

Speech signals were the Russian names of numbers from 1 to 10. Noises of fourteen kinds were used for speech signals noising (Table II). These were noises of household appliances and computers, street and transport, teaching rooms and lobbies.

Toolkit HTK was used for ASR system simulation and recognition accuracy assessment [7]. There were 22 phonemes of Russian language in phoneme vocabulary and there has been used 39 MFCC_0_D_A coefficients when ASR simulating. Clean speech signals (single words) were recorded in anechoic room ($T_{60} \approx 0.1$ s). Parameters of digitized sounds were: sampling rate 22050 Hz, linear quantization 16 bit. Signal-to-noise ratio (SNR) was near 45 dB for saved "clean" speech signals. Every word of clean speech was recorded 20 times; the words were uttered by speaker-woman with a different intonation.

Testing of ASR system was performed on six samples of noisy speech. Test sentences consisted of all ten words, with pauses between them 0.3–0.5 s. The recognition accuracy

$$Acc\% = \frac{N - D - S - I}{N} \times 100\%$$

was assessed according to the test results, where N is the total number of labels in the reference transcriptions; D is the number of deletion errors; S is the number of substitution errors; I is the number of insertion errors.

III. Experimental Results

Test results of ASR system trained on clean speech are shown in Table II and Fig. 1. These results indicate that the quality of recognition depends essentially on the spectral and temporal properties of noise. Indeed, for speech in trolley, $Acc\%$ = 95% for SNR > 17 dB, and for speech masked by noise of people filled audience, $Acc\%$ = 95% for SNR > 25 dB. Noise in the underpass has the most powerful masking properties. This can be explained both the combined action of noise and reverberation, and spectral-temporal characteristics of the noise [8] – [10].

Test results of ASR system trained on noised speech by FMT technique are shown in Fig. 2. Transport noise of street paved with stone blocks was used here. These charts are in good agreement with the results of [3], and they also give a more comprehensive picture of the FMT technique. Indeed, as follows from Fig. 1, when training on clean speech, recognition accuracy $Acc\%$ = 95% for paved street noise is achieved only for $SNR_r > 28$ dB. Meanwhile, when training on noised speech by FMT technique with $SNR_t = 10$ dB, recognition accuracy $Acc\%$ = 95% is achieved at $SNR_r = 7...15$ dB. When increasing SNR_t to 15 dB, it can be achieved $Acc\%$ = 95% for $SNR_r = 8...27$ dB. A further increasing of SNR_t to 20 dB provides $Acc\%$ =95% for $SNR_r = 12...35$ dB. As it can be seen, growth of SNR_t leads to expansion and shifting to the right values range of SNR_r that ensure the required accuracy of recognition.

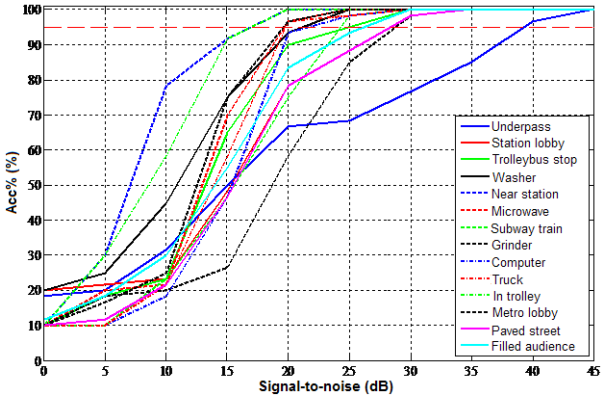

Fig. 1. Acc% for ASR system trained on clean speech.

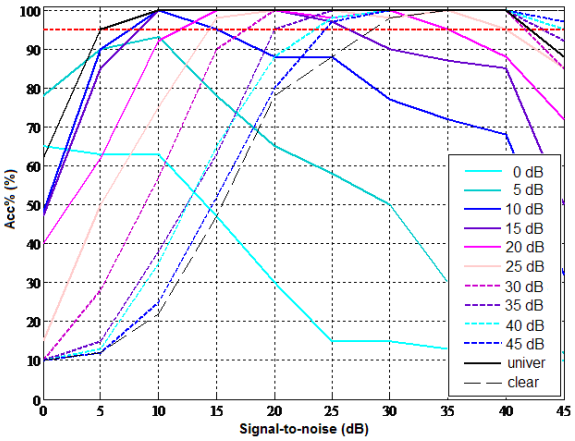

Fig. 2. Acc% for ASR system trained by FMT technique.

Test results of ASR system trained on noised speech by SMT technique are shown in Table III. It can be seen that this training method is much superior to FMT technique. For noise of paved street, recognition accuracy $Acc\%$ = 95% is achieved for $SNR_r \geq 5$ dB, and for most other types of noise the same accuracy is achieved for $SNR_r \geq 10$ dB. The exceptions are the subway train noise and noise in the people filled auditorium – in these cases, the recognition accuracy $Acc\%$ =95% is achieved only for $SNR_r \geq 25$ dB.

TABLE II. ACC% VALUES FOR CLEAN SPEECH TRAINING

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

TABLE III. ACC% FOR NOISED SPEECH TRAINING BY MEANS OF SMT TECHNIQUE

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

IV. CONCLUSION

Several techniques of ASR system training have been compared.

They are technique of training on clean speech signals, and also two techniques, FMT and SMT, of training on noised speech signal.

It is shown, for eleven of the fourteen kinds of noise interference, that SMT technique allows to reach the 95% recognition accuracy for $SNR_r \geq 10$ dB. When using FMT technique, it is also possible to achieve high recognition accuracy, but the technique is much more demanding to the volume of ASR system memory. When training on clean speech, 95% recognition accuracy was reached only for five of the fourteen kinds of noise interference for $SNR_r \geq 20$ dB. Thus, the degree of SMT training technique superiority over competing methods was experimentally assessed.

It would be useful in further to compare MT and SMT techniques with usage of the same kinds of noise interference.

REFERENCES

[1] Researchers fine-tune F-35 pilot-aircraft speech system. Available: https://web.archive.org/web/20071020030310/http://www.af.mil/news/story.asp?id=123071861

[2] E. Craparo, and E. Feron, “Natural Language Processing in the Control of Unmanned Aerial Vehicles”, Proceeding of AIAA Guidance, Navigation, and Control Conference, pp. 1-13, August 2004.

[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] 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.

[5] 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.

[6] 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.

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

[8] 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.

[9] 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.

[10] K. Pylypenko and 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.