Journal of Applied Science and Engineering Innovation Vol.2 No.2 2015
ISSN (Print): 2331-9062
ISSN (Online): 2331-9070

Subjective Evaluation of Speech Recognition in Noise

Jingfei Yang[1], Zhiling Hong[2]

[1] *Department of Computer Science, Xiamen University, Xiamen 361005, China*
[2] *Software School, Xiamen University, Xiamen 361005, China*

Abstract*:* In order to solve the current difficulties of modeling speech communication ability in Automatic Speech Recognition System (ASR), a new modeling method based on subjective assessments is proposed for the intelligent objective evaluation. And the data analysis and design procedure are provided in the same time. Finally, the experimental results have shown efficiency and rationality of this new method.

Keywords Computation; Speech Recognition; Noise Environment; Experimental Paradigms

INTRODUCTION

Speech is the extremely important method of communication between humans. In recent years, more and more people have researched in this area including automatic speech recognition. Intelligent speech recognition system is mainly used in the fields that help people detecting speech successfully. However, the transmission of speech can be affected by the background noise and distortions in the communication devices [1]. The speech recognition modeling system is classical information system and complex system, so its core problem is the speech recognition modeling. So far, there are still no efficient modeling methods for it because of the complexity of applications. It is important to find experimental paradigms to build an effective computational method for evaluating speech recognition in the noisy environment.

Since processing of noise in the normal environment interferes with processing of target speech leading to impaired processing of target speech. As features derived from speech have proven to be the most effective in automatic systems, in order to automatically extract information transmitted in speech signal, figuring out how the auditory system processes speech in noisy environment is crucial. In this study, the effect of speech recognition under the noisy environment was investigated. The speech corpus was generated by the speech-synthesis method.

MATERIALS AND METHODS

Materials

Speech synthesis method is provided by Shuyang Cao in 2011[Cao *et al.*, 2011], which introduce HMM-based speech synthesis method from qualitative Chinese speech corpus to quantitative.

The acoustic analog outputs were delivered to a loudspeaker in the central front of the participant. Speech stimuli were Chinese sentences and each of the sentences has 6 words including three key components. In Mandarin Chinese, a large number of words are two-character compound words in which each of the syllables has its own semantic representation. Listener may have to access the meanings of both syllables in order to get the whole word correct.

In this study, based on the database of the Chinese newspaper, the double-syllable verbs which were rated as having high frequencies of occurrence, and the double-syllable nouns which were also rated as having high frequencies of occurrence were used.

Both speech corpuses were developed by artificially synthesized young-female voices, and acoustic signals of target and masker speech for each of the three target young-female voices were generated by the Hidden Markov Model based speech-synthesis system.

In every-day speech communication situations, the amount of speech masking is highly dependent on the similarity of the target and masker voices.

Since the sequential dependences between the syllables in the two-syllable Chinese words could also affect the degree of noise masking, then access to the meaning of the first syllable might be expected to facilitate access to the meaning of the second syllable. Hence, the manner in which meaning of words is accessed may differ substantially across languages and influence the nature of informational masking in these languages.

In this study, about 1000 sentences from the Chinese nonsense sentences database was used for training. Speech signal was sampled at 16KHz, windowed by a 25-ms Blackman window with a 5-ms shift, and then mel-cepstral coefficients were obtained by the mel-cepstral analysis technique [Yoshimura, T. *et al.*, 1999]. A 5-state left-to-right HMMs with single diagonal Gaussian output a

Corresponding Author: Zhiling Hong, Software School, Xiamen University, Xiamen 361005, China.

distribution was adapted. By considering relationship between static and dynamic features during parameter generation, smooth spectral sequences are generated according to the statistics of static and dynamic parameters modeled by HMMs, resulting in natural sounding speech [Yoshimura, T. *et al.*, 1999].

The noise masker was a stream of steady-state speech-spectrum noise. The speech masker was a 47-s loop of digitally-combined continuous recordings for Chinese nonsense sentences.

Both noise speech and target speech sounds were presented at a level of 60 dBA. The sound pressure levels of maskers were adjusted to produce four signal-to-noise ratios (SNRs): -8, -4, 0, and 4 dB.

Design and Procedure.

There were three types of speech recognition conditions, and about eighteen Mandarin-speaking young university students participated in this study. All the participants had normal and bilaterally balanced pure-tone-hearing thresholds.

For a testing session, participants were informed of both the masking condition and the noise type. Each trial was started with the noise phase. Participants' task was to determine which stimulus was the target sentence and the performance for each participant was scored.

Since knowledge of all but the last word could not be used to predict the last word. Because the targets were all nonsense sentences, and such knowledge could, however, help the listener identify and track the target sentence in the noise masker.

28 whole-course target sentences out of 56 in a testing session were recited by each of the three target voices. About 300 priming sentences and 700 target sentences were used in this study.

Information quantity of a sentence was the sum of information quantities of the key words. All the lists of nonsense sentences were constructed in such a way that the information quantity of each list was about the same.

RESULTS AND DISCUSSION

As mentioned above, each of the key words in the nonsense sentence had two syllables. Because access to the meaning of the first syllable might affect the identification of the second syllable, we determined the number of times the first syllable was correctly identified, the number of times the second syllable was correctly identified, and the number of times both of the syllables in the word were correctly identified.

A logistic psychometric function was fitted to each participant's data, the results showed that there was a significant effect of masker type ($p = 0.000<0.005$) and a significant main effect of target sentence stimuli condition ($p=0.000<0.005$), but no significant interaction between noise masker type and target sentence condition ($p>0.01$).

Multiple t-tests (Bonferroni corrected) indicated that all noise type conditions differed significantly from one another. Hence, thresholds were lower for the noise type was two-people-speaking than the noise type was steady white noise, indicating that the target sentence provided a release from masking.

In addition, A two-factor, within-subject ANOVA confirmed that there was a significant main effect of masker type ($p=0.000<0.05$), a significant main effect of noise condition ($p=0.000 <0.05$), but no significant masker by noise condition interaction ($p = 0.129>0.1$).

Multiple t-tests (Bonferroni corrected) indicated that the slopes were shallower when the masker was two-people-speaking speech than when it was steady-white-noise. Shallower slopes in the presence of a noise, coupled with an asymptote of 100% correct at the higher SNRs under all conditions means that the amount of release from noise masking increased as SNR decreased. Hence, the lower the SNR, the greater was the amount of release from masking due to the presentation of a target speech.

The differences in threshold μ between conditions were also examined. A one-way Analysis of variance confirms that the effect of noise type was not significant ($p = 0.461>0.05$). However, when the masker was two-people speaking speech, the performance under the steady-white noise condition was much poorer than those under other noise conditions. A one-way ANOVA shows that the effect of noise type was significant ($p = 0.003<0.01$). Post hoc analyses show that the threshold under the two-people speaking noise condition was significantly better than that noise masking condition ($p = 0.003<0.01$).

It is important to note that the effect of a different noise type did not depend on the order in which conditions were experienced.

It would have expected such order effects if the steady-white noise type exerted its effect primarily by familiarizing the listener with the speech and noise characteristics.

In this study's case, the listener would have no exposure prior to experiencing the noise masking condition and therefore might be expected to show a larger release from masking than in the latter case where the amount of exposure to the noise would be extensive before the speech masking condition was experienced.

CONCLUSION

In this paper, we want to seek for system analysis model for speech recognition. From this analysis process and experiment, we endeavor to seek some common subjective methods of building intelligent automatic speech recognition system. And in the course of this exploration, we are able to see that comprehensive integration is inevitable trend in designing intelligent robotic system.

The results of this study show that under each of the stimulus conditions, percent-correct word scores

increased monotonically with the increase of SNR from -8 dB to 4 dB, without displaying plateaus.

Any manipulation, as long as it helps distinguish the target and direct selective attention toward the target, will help segregate target speech from competing noise.

Indeed, a comparison of deep troughs of envelopes between the Chinese two-talker speech masker used in the present study indicates that there appears to be a greater degree of amplitude modulation in the Chinese speech envelope. It is important to note, however, that a number of factors, such as speech rate, will affect the frequency and depth of troughs in a language

It is well known that subjective assessment can provide strong basis for objective assessment and this subjective computational method can provide a useful method for improving the automatic speech recognition systems

ACKNOWLEDGMENT

The authors wish to thank the helpful comments and suggestions from my teachers and colleagues in speech and hearing research center. This work is supported by the 2014 Program for New Century Excellent Talents in Fujian Province University, and the Open Funding Project of Zhejiang Key Laboratory for Research in Assessment of Cognitive Impairments.

REFERENCES

Bronkhorst, A. W. The cocktail party phenomenon: A review of research on speech intelligibility in multiple-talker conditions. Acustica,86(1),pp.117-128,2000.

Gong, Y. Speech recognition in noisy environments: A survey. Speech communication, 16(3), pp.261-291,1995.

Cherry, E. C. Some experiments on the recognition of speech, with one and with two ears, J. Acoust. Soc. Am. 25, pp.975-979,1953.

Freyman, R. L., Balakrishnan, U., and Helfer, K. S. Effect of number of masking talkers and auditory priming on informational masking in speech recognition, J. Acoust. Soc. Am. 115, pp.2246-2256, 2004.

Yoshimura, T., Tokuda, K., Masuko, T., Kobayashi, T., and Kitamura, T. Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis, Proc. EUROSPEECH, 5, pp.2347-2350,1999.

Tokuda, Keiichi, Heiga Zen, and Alan W. Black. An HMM-based speech synthesis system applied to English. Speech Synthesis, pp. 11-13, 2002.

Wolfram, S. Mathematica: A System for Doing Mathematics by Computer. Addison-Welsey, New York , 1991.

Wu, M.-H., Li, H.-H.,Hong, Z.-L., Xian, X.-C., Li, J.-Y., Wu, X.-H., Li, L. Effects of aging on the ability to benefit from prior knowledge of message content in masked speech recognition. Speech Communication, 54,529-542,2012.

Yang, Z. G., Chen, J., Wu, X. H., Wu, Y. H., Schneider, B. A., and Li, L. The effect of voice cuing on releasing Chinese speech from informational masking, Speech Communication, 49, pp. 892-904,2007.

Cao, S.-Y., Li, L., and Wu, X.-H. Improvement of intelligibility of ideal binary-masked noisy speech by adding background noise. Journal of the Acoustical Society of America, 129, pp.2227-2236, 2011.

Masuko, T., Tokuda, K., Kobayashi, T., and Imai, S. Speech synthesis from HMMs using dynamic features, in Proceedings of the International Conference on Acoustics, Speech, and Signal Processing, ICASSP-96, Atlanta, GA, pp.389–392,1996.

Krishnan, A., Xu, Y.S., Gandour, J., Cariani, P., 2005. Encoding of pitch in the human brainstem is sensitive to language experience. Cog. Brain Res. 25, 161–168.

Assmann, P.F., Summerfield, Q., 1989. Modeling the perception of concurrent vowels – vowels with the same fundamental-frequency. J. Acoust. Soc. Amer. 85, 327–338

Festen, J.M., Plomp, R., 1990. Effects of fluctuating noise and interfering speech on the speech–reception threshold for impaired and normal hearing. J. Acoust. Soc. Amer. 88, 1725–1736.

Schneider, B.A., Li, L., Daneman, M., 2008. How competing speech interferes with speech comprehension in everyday listening situations.

Summers, V., Molis, M.R., 2004. Speech recognition in fluctuating and continuous maskers: effects of hearing loss and presentation level. J.Sp. Lan. Hear. Res. 47, 245–256.J. Amer. Acad. Audiol.

Wu, X.-H., Wang, C., Chen, J., Qu, H.-W., Li, W.-R., Wu, Y.-H.,Schneider, B.A., Li, L., 2005. The effect of perceived spatial separation on informational masking of Chinese speech. Hear. Res. 199, 1–10.