Analysis of HMM-Based Lombard Speech Synthesis

Tuomo Raitio[1], Antti Suni[2], Martti Vainio[2], and Paavo Alku[1]

[1]Aalto University, Department of Signal Processing and Acoustics, Espoo, Finland
[2]University of Helsinki, Department of Speech Sciences, Helsinki, Finland

Aalto University
School of Electrical
Engineering

UNIVERSITY OF HELSINKI

Introduction

- Humans modify their voice in interfering noise in order to maintain the intelligibility of their speech – this is called the Lombard effect
  → increased loudness, modified spectral qualities, durations, and prosody
- Lombard speech is less studied in speech synthesis, but hidden Markov model (HMM) based speech synthesis provides good opportunity for this
- HMM-based synthesizer GlottHMM was used as a platform for this study due to the flexibility of the vocoder
- We have studied three methods for generating synthetic Lombard speech:

1. Vocoder Modification

- Vocoder of the synthesizer and speech parameters are modified to generate Lombard effect:
  → Rate of speech lowered, pitch raised and pitch range compressed
  → Spectral tilt was decreased in order to concentrate more energy to formant frequencies
  → Stronger postfiltering was applied to generate a more prominent formant structure
  → Speech signal was companded to increase loudness
- Most intelligible of all systems in the Blizzard Challenge 2010 speech-in-noise task

2. Adaptation

- Three hundred sentences were recorded, while 83 dB babble noise was played to speaker's ears through headphones
- Speaker's voice was fed back to headphones to control the degree of the Lombard effect
- Lombard sentences were used to adapt a 600-sentence base voice with CSMAPLR + MAP method
- Adaptation applied to all streams using state-tying decision trees for regression classes

3. Extrapolation

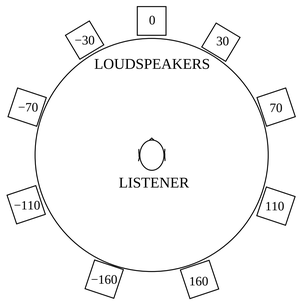
- Interpolating and extrapolating between two or more model sets is a unique feature in HMM-based TTS
- Extrapolation ratio between the normal voice and the adapted Lombard voice was set to 1.5 (where 0 corresponds to the normal voice and 1.0 to the adapted voice)
- Extrapolation applied to all streams except for duration where adapted models were used

Listening Tests – Setup

- Subjective listening tests were conducted in a standardized listening room (ITU-R BS.1116-1)
- Realistic noise environment was created by playing a real multichannel recording of street noise from nine identical large speakers (Genelec 8060A) (see figure below), speech was played through the center speaker
- Three noise levels were selected: silence, moderate noise (63 dB), and extreme noise (70 dB)
- Speech stimuli: phonetically and lexically balanced short sentences
- The loudness of speech samples was normalized by ITU-T P.56
- Average SNRs were −1 dB and −8 dB for moderate and extreme noises, respectively
- 17 persons performed both intelligibility test and subjective evaluation, 90 sentences each

Test voices and their averaged A-weighted SPLs:

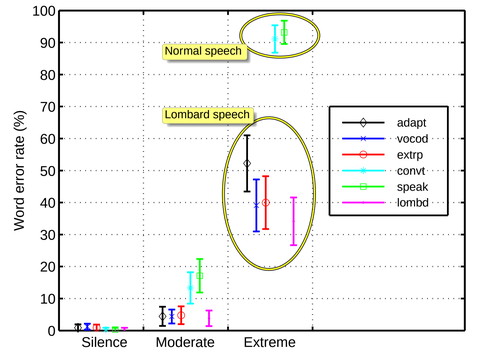
Voice	Description	SPL
adapt	Lombard synthesis by adaptation	62 dB
vocod	Lombard synthesis by modification of the vocoder	63 dB
extrp	Lombard synthesis by extrapolation	63 dB
convt	Normal speaking style synthesis	61 dB
speak	Natural normal speaking style speech	59 dB
lombd	Natural Lombard speech	63 dB



Listening test setup

Listening Tests – Intelligibility

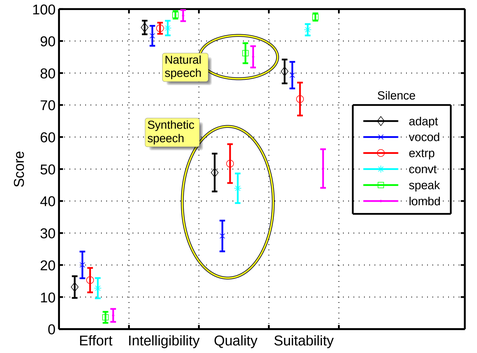
- First, an intelligibility test was performed, where short Finnish sentences were presented to the listener
- The listeners were allowed to listen to the samples only once, and were then asked to type in what they heard
- Word error rates were evaluated taking separately into account the inflectional and derivational suffixes
- In silence, all voice types are equally intelligible
- In moderate noise, the WERs of all Lombard voices (*adapt*, *vocod*, *extrp*, *lombd*) show no statistical difference, whereas both normal speaking style voices (*convt*, *speak*) are statistically less intelligible
- In extreme noise, normal speech (*speak) and synthesis (convt) are almost totally unintelligible, while the WER's of two synthetic Lombard voices (vocod, extrp) show no statistical difference to natural Lombard speech (lombd)*

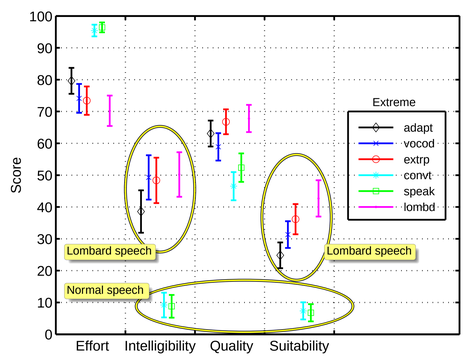


Results of the intelligibility test

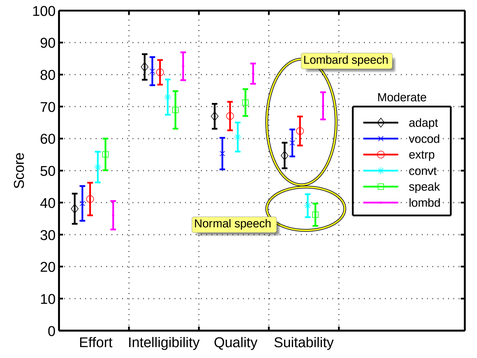
Listening Tests – Subjective Evaluation

- Second, the listeners were asked to rate the samples according to four questions (effort, intelligibility, quality, suitability)
- The listeners could listen to the samples as many times as desired
- In silence, both natural voices (*speak, lombd*) are rated higher in quality compared to synthetic ones
- The suitability of all Lombard voices are rated relatively low in silence
- In noise conditions, the suitability scores are completely opposite, Lombard voices being more appropriate than conventional voices
- All the Lombard voices are rated better than conventional speaking style voices in terms of intelligibility and required effort





Silence →
Moderate noise →
← Extreme noise



Conclusions

- The study shows that the synthesized Lombard voices are more intelligible and more suitable to be used in the presence of noise compared to conventional voices
- Synthesized Lombard voices are rated to be similar to natural Lombard speech
- The adapted Lombard voices are considered of higher quality than the Lombard voice generated by vocoder modification, but the latter can be useful if no adaptation data is available

tuomo.raitio@aalto.fi, antti.suni@helsinki.fi