Expressão de Emoção na Voz Gerada por Conversão Texto-Fala

Marcos E. Crivellaro[1], Rogério E. da Silva[1]

[1]Departamento de Ciências da Computação – Universidade do Estado de Santa Catarina (UDESC)
Centro de Ciências Tecnológicas (CCT) – Joinville – SC – Brasil
markitou@gmail.com, rsilva@joinville.udesc.br

Abstract. *This paper studies tools involved in the text-to-speech area of speech synthesis. Through existing tools, speech samples were generated aiming the evaluation of naturalness and intelligibility of artificial voice by questionnaire, with a focus on emotional expression in speech. The results show that the samples have good quality for intelligibility, but medium quality for naturalness and bad quality for emotional expression. The next step of this work is to try to attach emotional characteristics in voice, to be informed by constructing phrases in specific script. The new results will be evaluated and compared to the initial results.*

Resumo. *Este trabalho estuda ferramentas envolvidas na área de conversão-texto fala da síntese de voz. Através de ferramentas já existentes, foram geradas amostras de fala com intuito de avaliar através de questionário as características de naturalidade e inteligibilidade da voz artificial, com foco na expressividade emocional na fala. Os resultados indicaram que em questão de inteligibilidade as amostras demonstram boa qualidade, mas qualidade mediana quanto à naturalidade da voz e péssima quanto à expressão emocional. O próximo passo deste trabalho é tentar atribuir características emocionais na voz, a serem informadas montando a frase em script específico. Os novos resultados serão avaliados e comparados aos resultados iniciais.*

1. Introdução

A síntese de voz (processo de produção artificial da voz humana) é uma tecnologia que vem apresentando grandes avanços. Contudo, o resultado da geração de fala digitalizada ainda necessita buscar melhorias, já que os sintetizadores de voz disponíveis ainda não são capazes de reduzir a artificialidade da voz a um nível imperceptível. Isto permite aos ouvintes perceber que se trata de uma máquina, e não um ser humano que está falando.

2. Estado da Arte

O ser humano vive em sociedade, tornando a comunicação indispensável no dia-a-dia. Cada passo dado por um indivíduo reflete em seu futuro em seu ambiente social.

Sinais emocionais emitidos por um indivíduo transmissor podem ser interpretados de diversas formas por seu observador. A expressão vocal é reconhecida como uma das principais características da voz, e também um dos principais sinais afetivos.

3. Desenvolvimento

Este trabalho propõe o desenvolvimento de uma ferramenta com capacidade de encenar scripts em formato específico, contendo frases, expressões e detalhes relevantes para que a sintetização realizada pelo software se aproxime da fala natural.

Primeiramente foi feita uma avaliação de uma ferramenta já existente, disponibilizando um questionário com 23 perguntas em uma página da web, organizadas numericamente e agrupadas em páginas de acordo com as amostras de som a que se referiam, sendo que estas amostras continham frases semelhantes, sintetizadas com diferentes *engines* de voz da Microsoft SAPI. O objetivo do avaliador ao responder o questionário era identificar traços de naturalidade, inteligibilidade e expressão emocional nas vozes sintetizadas, podendo optar por valores de 1 a 6 para avaliar a característica referida na pergunta. Houveram no total 137 respostas ao questionário.

Os resultados obtidos se apresentaram dentro do quadro esperado, onde a hipótese era de que o nível de expressividade emocional nas amostras de uma ferramenta já existente apresentaria resultados de baixa escala (péssimos, valores 1 e 2 na maioria dos casos). Em comparação, a naturalidade de forma geral foi avaliada como mediana (3 e 4), enquanto a inteligibilidade apresentou os melhores resultados (a maioria 5, com alguns 4).

4. Conclusões

Os resultados obtidos demonstraram a necessidade de melhor tratar as informações textuais, de forma que se possa obter mais dados da expressividade emocional e sintetizar a voz de forma adequada.

Para a próxima etapa, parâmetros de emoção na voz serão levantados para desenvolver uma ferramenta onde as frases serão montadas e inseridas na forma de script XML, apresentando parâmetros como velocidade de fala, emoção, entre outros, que buscam informar o modo como o locutor está falando. Dependendo destes parâmetros, a ferramenta manipulará as configurações do sintetizador de forma a tentar se aproximar o melhor possível da emoção desejada.

Referências

HOFER, G. O. “Emotional Speech Synthesis”. 2011. 179 p. Dissertação (Master of Science). School of Informatics, University of Edinburgh, Edinburgh. 2004. Disponível em <https://www.inf.ed.ac.uk/publications/thesis/online/IM040151.pdf>. Acesso em: 18 dez. 2011.

SCHRÖDER, M. "Emotional Speech Synthesis: A Review". In: DFKI, Saarbrücken, 2001. Institute of Phonectics, University of Saarland, Saarland, 2001. Disponível em <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.1.9050&rep=rep1&type=pdf>. Acesso em: 18 dez. 2011.

SILVA, D. da C. “ALGORITMOS DE PROCESSAMENTO DA LINGUAGEM E SÍNTESE DE VOZ COM EMOÇÕES APLICADOS A UM CONVERSOR TEXTO-FALA BASEADO EM HMM”. 2011. 179 p. Disponível em <http://www.pee.ufrj.br/teses/textocompleto/2011033102.pdf>. Acesso em: 18 dez. 2011.