論 文 内 容 の 要 旨

博士論文題目　Emphasis Speech-to-Speech Translation Considering Acoustic and Linguistic Features

（邦題：音響・言語特徴を考慮した強調音声翻訳）

氏　　名　Do Quoc Truong

要　　旨

Speech-to-speech translation (S2ST) systems are capable of breaking language barriers in crosslingual communication by translating speech across languages. However, there are still problems remaining. First, existing emphasis modeling techniques assume emphasis speech is expressed at word-level with binary values indicating the change of acoustic feature. However, depending on the context and situation, emphasis can be expressed at arbitrary levels. This assumption also limit the capability of the model in the way that it can only generate binary emphasized speech. Second, the existing emphasis S2ST approaches used for emphasis translation is not optimal for sequence translation tasks and cannot easily handle the long-term dependencies of words and emphasis levels. Third, the whole translation pipeline still separates emphasis and standard S2ST systems, making it not possible to perform joint optimization and inference. And finally, only binary levels of acoustic feature (emphasis speech) is taken into account while emphasis can be expressed in many ways including written form at arbitrary levels. This problem limits the capable of emphasis S2ST system that it can only translate acoustic features but not linguistic features of emphasis. This thesis attempts to solve the problems above by (a) proposing an approach that can handle continuous emphasis levels in both emphasis modeling and translation components, and (b), combining machine and emphasis translation into a single model, which greatly simplifies the translation pipeline and make it easier to perform joint optimization. And finally, (c), we propose a data-driven approach on studying correlation of emphasis expressed in both text and speech as a first step toward

acoustic-linguistic emphasis translation. With regards to the experiments, the results on continuous emphasis modeling demonstrated its effectiveness in a emphasis detection task while producing more natural synthetic speech. Experiments on an emphasis translation task utilizing sequence-to-sequence approach with continuous emphasis levels show a significant improvement over previous models in both objective and subjective tests. Moreover, the evaluation on joint translation model also show that our models can jointly translate words and emphasis with one-word delays instead of full-sentence delays while preserving the translation performance of both tasks. Finally, our studies on emphasis representations in both audio and text forms have investigated the way humans express emphasis in different contexts and analyzed ambiguities between emphasis levels.

氏　名	Do Quoc Truong

（論文審査結果の要旨）

Speech-to-speech translation (S2ST) systems are capable of breaking language barriers in crosslingual communication by translating speech across languages. However, there are still problems remaining. First, existing emphasis modeling techniques assume emphasis speech is expressed at word-level with binary values indicating the change of acoustic feature. However, depending on the context and situation, emphasis can be expressed at arbitrary levels. This assumption also limit the capability of the model in the way that it can only generate binary emphasized speech. Second, the existing emphasis S2ST approaches used for emphasis translation is not optimal for sequence translation tasks and cannot easily handle the long-term dependencies of words and emphasis levels. Third, the whole translation pipeline still separates emphasis and standard S2ST systems, making it not possible to perform joint optimization and inference. And finally, only binary levels of acoustic feature (emphasis speech) is taken into account while emphasis can be expressed in many ways including written form at arbitrary levels. This problem limits the capable of emphasis S2ST system that it can only translate acoustic features but not linguistic features of emphasis. This thesis attempts to solve the problems above by (a) proposing an approach that can handle continuous emphasis levels in both emphasis modeling and translation components, and (b), combining machine and emphasis translation into a single model, which greatly simplifies the translation pipeline and make it easier to perform joint optimization. And finally, (c), we propose a data-driven approach on studying correlation of emphasis expressed in both text and speech as a first step toward acoustic-linguistic emphasis translation.

The research proposed solutions to the problems which haven't been solved and series of his research resulted in two journal papers and eight peer reviewed international conference papers. As a result, the thesis is sufficiently qualified as Doctoral thesis of Engineering.