論 文 内 容 の 要 旨

博士論文題目　Multimodal Machine Chain
（深層学習に基づくマルチモーダルチェインモデルに関する研究）

氏　　名　Johanes Effendi The

（論文内容の要旨）

Researchers have been working in speech technology for many decades. State-of-the-art automatic speech recognition (ASR) and text-to-speech synthesis (TTS) systems are currently based on end-to end deep learning frameworks. Traditionally, they are usually trained by applying supervised learning techniques that rely on the availability of parallel speech data and its corresponding transcriptions. To improve the performance in the presence of unexpected acoustic variability, we usually collect more data to train more detailed models. Unfortunately, such a method can only be used to train the model for about 10-20 of the world's most common languages. For many others, the parallel data of speech and its transcriptions are usually unavailable, which makes such models hard to implement.

On the other hand, human learning does not rely on parallel data. We can learn from any experience, even if the examples are not provided at the same time. These experiences are perceived in the form of senses, such as auditory and visual, which shares complementary behaviour to ensure flexible learning from any modality (i.e. speech, text, image) in the form of a feedback loop. Inspired by this mechanism, we propose a multimodal machine chain (MMC) as a general framework that accommodates learning in any kind of modality and data availability (i.e. paired, unpaired, single modality). In this framework, a cross-modal model is able to learn from non-parallel data through feedback it receives after mapping the input into other modalities. Consequently, more modalities, in this case, means more feedback can be made, which therefore enable model learning with fewer data. This makes our proposed learning strategy beneficial for under-resource language, where such technologies matter the most.

This thesis contribution is four-fold. First, we defined a general framework that enables

氏　名	Johanes Effendi The

(論文審査結果の要旨)

State-of-the-art automatic speech recognition (ASR) and text-to-speech synthesis (TTS) systems are currently based on end-to-end deep learning frameworks. Traditionally, they are usually trained by applying supervised learning techniques that rely on the availability of parallel speech data and its corresponding transcriptions. On the other hand, human learning does not rely on parallel data. Humans learn from any experiences that are perceived in the form of senses, such as auditory and visual, which share complementary behavior to ensure flexible learning from any modality (i.e., speech, text, image) in the form of a feedback loop. Inspired by this mechanism, this thesis proposes a multimodal machine chain (MMC) as a general framework that accommodates learning in any modality and data availability (i.e., paired, unpaired, single-modality). This thesis contribution is four-fold.

1) Thesis defined a general framework that enables cross-modal model training in any modality and any data availability.

2) Thesis showed that the proposed MMC framework can be used to enable semi-supervised cross-modal collaboration that allows learning from a single-modality data, which modality is unrelated.

3) Thesis pushed the level of supervision boundary into weakly-supervised learning, to enable a speech-to-text mapping using a visually-connected non-parallel data.

4) Thesis showed our proposed MMC framework capability to learn a self-supervised discrete speech representation to enable image-to-speech generation without text.

All these four contributions in the form of the MMC framework and its applications show its capability to enable speech processing model learning.

The proposed MMC provides a general framework and can be applied to the various application. The thesis research brought a very new framework and evidenced the novelty with various experiments. A series of his research resulted in three high-quality peer-reviewed journal papers, three peer-reviewed international conference papers. As a result, the thesis is sufficiently qualified as a Doctoral thesis of Engineering.