論 文 内 容 の 要 旨

博士論文題目　Unsupervised Representation Learning and Acoustic Modeling in the Zero Resource Scenario

（邦題：ゼロリソース条件における音響単位と音響モデルの教師なし学習の研究）

氏　　名　Michael Heck

要　　旨

Automatic speech recognition (ASR) has experienced a remarkable development over the decades. Technological advances however have largely been made on a small subset of all human languages that are rich in resources. This has two main consequences: Methods for ASR evolved such that they learn best from massive amounts of data, and the majority of languages, spoken by billions of speakers, has been neglected for a long time. Moreover, most of the world's languages have no written form and are therefore severely under-resourced. If besides raw speech data no other information about a language is available, we speak of a zero resource scenario. Inferring models in such a scenario is a challenging task, which can be compartmentalized into unsupervised learning of lexical units and unsupervised subword modeling.

This thesis addresses the problem of unsupervised subword modeling in the zero resource scenario. The two major challenges of unsupervised subword modeling are representation learning and model design. Representation learning is the task to find speaker independent, robust speech features without prior knowledge that highlight linguistically relevant properties and suppress irrelevant information. Model design is the task to develop and infer a structure that approximates the true distributions of speech better than previous models.

This thesis approaches the representation learning problem by elaborating a novel framework for unsupervised subword modeling that takes advantage of automatically estimated feature transformations (Chapter 4). The proposed

algorithm jointly learns transformations for the speech input without prior category knowledge and infers a Dirichlet process mixture model (DPMM) that represents sound classes. The incorporation of feature transformations into the unsupervised subword modeling framework considerably supports finding speaker independent, robust representations with high class discrimination properties. The proposed method proved its effectiveness in actual performance evaluations and delivered state-of-the-art performance in the zero resource challenges 2015 and 2017. The construction of a functional acoustic unit tokenizer shows that the found acoustic units carry meaning which can be utilized to solver higher-level problems (Chapter 5).

The model design problem is addressed by the introduction of a novel design for a Dirichlet process mixture of mixtures model (Chapter 6). Speech is inherently complex and requires models of appropriate complexity for proper representation. A long standing assumption in ASR research is that the emission of speech representations is modeled by multimodal distributions. As opposed to the unimodal modeling assumption of a standard DPMM, the novel algorithm proposed in this thesis can infer a mixture of mixtures to discover clusters in raw data that are made up of multimodal distributions. In experiments, the proposed design leads to the inference of fewer classes that represent subword units more consistently and show longer durations, which is a first step towards a fully unsupervisedly learned model for speech that represents units of appropriate length and complexity.

The methods presented in this thesis are ultimately designed towards enabling low and zero resource automatic speech recognition and provide a good basis for further research on the possibilities of learning acoustic units and acoustic features from scratch, without any prior category knowledge or other meta information about the target language.

氏　名	Michael Heck

（論文審査結果の要旨）

This thesis addresses the problem of unsupervised subword modeling in the zero resource scenario. The two major challenges of unsupervised subword modeling are representation learning and model design. Representation learning is the task to find speaker independent, robust speech features without prior knowledge that highlight linguistically relevant properties and suppress irrelevant information. Model design is the task to develop and infer a structure that approximates the true distributions of speech better than previous models.

This thesis approaches the representation learning problem by elaborating a novel framework for unsupervised subword modeling that takes advantage of automatically estimated feature transformations (Chapter 4). The proposed algorithm jointly learns transformations for the speech input without prior category knowledge and infers a Dirichlet process mixture model (DPMM) that represents sound classes. The proposed method proved its effectiveness in actual performance evaluations and delivered state-of-the-art performance in the zero resource challenges 2015 and 2017. (Chapter 5). The model design problem is addressed by the introduction of a novel design for a Dirichlet process mixture of mixtures model (Chapter 6). As opposed to the unimodal modeling assumption of a standard DPMM, the novel algorithm proposed in this thesis can infer a mixture of mixtures to discover clusters in raw data that are made up of multimodal distributions. The methods presented in this thesis are ultimately designed towards enabling low and zero resource automatic speech recognition and provide a good basis for further research on the possibilities of learning acoustic units and acoustic features from scratch, without any prior category knowledge or other meta information about the target language.

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