Recent Advances in Speech Recognition for Spontaneous Speech Translation

Yoshinori Sagisaka

ATR Interpreting Telecommunications Research Laboratories, 2-2 Hikaridai Seika-cho Soraku-gun Kyoto, 619-0288

Abstract: This paper introduces recent advances in speech recognition research at ATR-ITL. The component recognition technologies are described from input speech signal processing to word candidate search using statistical linguistic constraints. The adoption of linguistically loose constraints makes it possible to recognize spontaneous speech whose linguistic structures are not definitely specified by a conventional grammar for written language. Component technologies form a set of flexible software tools ATRSPREC for building a recognition system and they are used in our new speech translation system MATRIX.

INPUT SPEECH SIGNAL PROCESSING

In most current recognition systems, conventional spectral parameters for speech coding are widely used to distinguish phonetic units. As these parameters are not designed for phonetic discrimination purposes, two types of parameters are designed for fine and robust recognition to cope with variations in speaking conditions and signal transmission environments. Dynamic cepstrum parameter (1) was proposed to incorporate human's perceptual characteristics. Through this parametrization, human's bi-directional masking characteristics was implemented in the form of cepstral liftering which has also been used with other techniques such as cepstral mean subtraction. The other type of parameter was looked for the best use of speech signal characteristics. Spectral sub-band centroid (SSC) parameter (2) was proposed to capture global spectral shapes. This parameter is very insensitive to noisy environments and showed higher performances when used with conventional parameters.

The other approach has been taken quite independently from a pure pattern classification viewpoint. The success of the generalized probabilistic descent (GPD) technique with the minimization of classification errors (MCE) derives the same methodology in feature extraction. A discriminative metric design scheme (3) has been proposed to automatically determine acoustic parameters specific to recognition targets by minimizing classification errors. This scheme incorporates feature extraction and discrimination of speech recognition units using the same MCE criteria.

ACOUSTIC MODELS AND THEIR TRAINING USING SPEECH CORPORA

For an acoustic model for speech recognition, automatic designing and training of HMM models is one of the key issues to build robust contextual dependent models with a small number of parameters. Maximum likelihood-based successive state splitting (ML-SSS) has been proposed to attain effective state-tying properties (4). This method enables exact factor (e.g. phonetic context) related tying independent of other factors encountered in training speech corpora (e.g. speaker variance). To feed back the recognition error characteristics to HMM state-tying, a parameter re-tying and retraining scheme (5) was also proposed.

To cope with the insufficiencies of HMM modeling, other acoustic models such as a segment model and a recurrent neural net model (6) have been studied to make better use of parameter dependencies between neighboring frames. In addition to these improvements of acoustic models by themselves, we also focus on stochastic pronunciation modeling to recognize spontaneous speech which has larger acoustic phonetic variations than read speech (7).

SPEAKER INDEPENDENT ACOUSTIC MODELING AND SPEAKER ADAPTATION

The variation of acoustic characteristics among speakers is one of the most difficult problems in speech recognition. To efficiently use the constraints due to speaker characteristics, a tree-structured speaker clustering technique has been proposed (8). In the training stage, starting from a speaker-independent model, speaker clusters are structured in a tree-format by measuring the similarities of acoustic characteristics between speakers. In the recognition stage, this tree is traced from top to bottom. The most adequate model is chosen in the trace through the comparison of likelihood among chosen models in all levels.

Speaker adaptation techniques have been studied to modify existing models to input speaker characteristics using small number of utterances. Maximum a posteriori estimation and transfer vector field smoothing (MAP-VFS) speaker adaptation was proposed to interpolate and smooth untrained or insufficiently trained model parameters by
statistically balancing the reliability of original parameters in existing models and newly estimated parameters using input utterances (9). Furthermore, to drastically reduce the speech corpora needed for adaptation, a new adaptation scheme is pursued based on the vocal-tract-size related parameters. On the other hand, an incremental Bayesian adaptation scheme (10) is proposed for a new user to use the system for a specific task after each utterance is recognized. The recognition system is adaptively trained only with the current utterance the user speaks, and updated models are used to recognize the next utterance.

LANGUAGE MODELING AND THE EXTRACTION OF LINGUISTIC CONSTRAINTS FROM LANGUAGE CORPORA

Word N-grams have been widely used as effective linguistic constraints to reduce search efforts in continuous speech recognition. A generation scheme of variable order N-grams (11) enabled to attain reliable statistical constraints from a given language corpus with fewer parameters than conventional N-grams and outperformed them. As with acoustic modeling, an adaptation scheme can be applied to statistical language models. As it takes huge efforts to collect language data to build N-gram models for each specific application, Maximum A Posteriori (MAP) adaptation was applied to N-gram construction (12). It has been experimentally confirmed that MAP adaptation works quite efficiently particularly when the target task data size is small.

EFFICIENT SEARCH OF WORD CANDIDATES AND SYSTEM INTEGRATION EFFORTS

In continuous speech recognition, especially fast search for real-time implementation, reducing the number of word (sentence) hypotheses is a crucial issue. An efficient search algorithm was proposed to generate word graphs with multi-pass search (13). In this algorithm, two likelihood-approximation, cross-word context approximation and lenient language score smearing were employed to reduce the computational cost for word graph generation. An experimental speech recognition system using this search shows real-time performance for about two thousand words and is used as a search engine for our speech translation system MATRIX.

To facilitate system development, current software technology has been efficiently employed in our speech recognition tool kit ATR-SPREC (14). Server-client architecture is adopted to combine calculation-intensive modules and user-interface oriented modules. Event driven processing is employed for each processing step to allow real-time operation. This architecture allows customized implementation of speech recognition systems with a minimal programming effort and fits well standard graphic user interface programming techniques.

REFERENCES


614