Annotating Speech Data for Pronunciation Variation Modelling

Per-Anders Jande
KTH: Department of Speech, Music and Hearing/CTT – Centre for Speech Technology

Abstract

This paper describes methods for annotating recorded speech with information hypothesised to be important for the pronunciation of words in discourse context. Annotation is structured into six hierarchically ordered tiers, each tier corresponding to a segmentally defined linguistic unit. Automatic methods are used to segment and annotate the respective annotation tiers. Decision tree models trained on annotation from elicited monologue showed a phoneme error rate of 9.91%, corresponding to a 55.25% error reduction compared to using a canonical pronunciation representation from a lexicon for estimating the phonetic realisation.

Introduction

The pronunciation of a word depends on the context in which the word is uttered. A model of pronunciation variation due to discourse context is interesting in a description of a language variety. Such a model can also be used to increase the naturalness of synthetic speech and to dynamically adapt synthetic speech to different areas of use and to different speaking styles.

The pronunciation of words in context is affected by many variables in conjunction. The amount of variables and their complex relations make data-driven methods appropriate for modelling. Data-driven methods are methods used to create general models from examples, e.g. using machine learning.

To use data-driven methods, data (examples) is of course a prerequisite. The method for acquiring data for variables hypothesised to be important for the pronunciation of words is to annotate recorded spoken language with information about the variables. The pronunciation and the set of context variables is thus used as an example, which can be used for finding general structures in the data.

This article describes methods for annotating speech data with information hypothesised to be important for predicting the segment-level pronunciation of words in discourse context.

Background

Work on pronunciation variation in Swedish on the phonological level has been reported by several authors, e.g. Gårding (1947), Bruce (1986), Bannert and Czigler (1999) and Jande (2003a, 2003b, 2004).

There is an extensive corpus of reports on research on the influence of different context variables on the pronunciation of words. Variables that have been found to influence the segmental realisation of words in context are foremost speech rate, word predictability (often estimated by global word frequency) and speaking style (cf. e.g. Fosler-Lussier and Morgan, 1999; Finke and Waibel, 1997; Jurafsky et al., 2001; van Bael, 2004).

Speech Data

The speech data used for pronunciation variation modelling has not been recorded specifically for this project, but has been collected from various sources. The speech corpus includes data recorded or made available for research within the fields of phonetics, phonology and speech technology in different earlier research projects. The speech data has been selected to be dialectally homogeneous, to avoid dialectal pronunciation variation. The language variety used is central standard Swedish.

The speech data has been recorded in different situations and speaking style related variables are defined from the speaking situation. The collection of speech data collected for the project includes radio news broadcast and interviews, spontaneous dialogues, elicited monologues, acted readings of children's books, neutral readings of fact literature and recordings of dialogue system interaction.

Methods and software for annotation has been developed using mainly the VAKOS corpus (Bannert and Czigler, 1999) as the target to be annotated. This corpus was originally recorded and annotated for the study of variation in consonant clusters in central standard Swedish. It consists of ~103 minutes of monologue from ten native speakers of central Standard Swedish.

Method

Automatic methods (with some minor exceptions) are used for annotation of spoken language data, where annotation is not supplied for the corpora used. The word level annotation is the base for all other annotation. The automatically obtained word boundaries and orthographic transcripts are manually corrected. In this way, relatively little work can give a large gain in annotation performance for most types of annotation.

The annotation system is built as a serialised set of modules, producing output at different levels. The output can be manually edited and used as input to modules later in the chain.

Annotation Structure

All annotation is connected to some duration-based unit at one of six hierarchically ordered tires. The tiers correspond to 1) the discourse, 2) the utterance, 3) the phrase, 4) the word, 5) the syllable and 6) the phoneme. Each tier is strictly sequentially segmented into its respective type of units. Some non-word units can be introduced in the word tier annotation to ensure that parts of the signal that are not speech can be annotated, e.g. pauses and inhalations.

A boundary on a higher tier is always also a boundary on a lower tier. An utterance boundary is thus also always a phrase boundary, a word boundary, a syllable boundary and a phoneme boundary. Thus, information can be unambiguously inherited from units on higher tiers to units on the tires below.

Having the information stored at different tiers enables easy access to the sequential context information, i.e., properties of the units adjacent to the current unit at the respective tiers.

Segmentation

Each annotation tier is segmented into its corresponding units, beginning at the word tier. Based on the word tier segmentation and information derived from the word units, the tiers above and below the word tier are segmented. The phoneme tier is segmented word-by-word using the orthographic annotation, a canonical pronunciation lexicon and an HMM phoneme aligner, NALIGN (Sjölander, 2003). The phonemes are clustered into syllables with forced syllable boundaries at word boundaries and the syllable tier is segmented using this clustering and the durational boundaries from the phoneme segmentation. Utterance boundaries are located manually with support from the word boundary annotation. The phrase tier is segmented utterance-by-utterance using the output of the TNT part of speech tagger (Brants, 2000; Megyesi, 2002a) and the SPARK parser (Aycock, 1998) with a context-free grammar for Swedish (Megyesi, 2002b).

Discourse Tier Annotation

The discourse annotation is related to speaking style characteristics and global speech rate. The speaking style/speaking situation variables included in the annotation are the *number of discourse participants* (monologue, two-part dialogue or multi-part dialogue), *degree of formality* (formal, informal), *degree of spontaneity* (spontaneous, elicited, scripted, acted, read), *discourse type* (human-directed, computer-directed). These variables are manually defined.

A number of automatically estimated measures of the average speech rate over the dialogue are also included. Speech rate is estimated by inverse segment duration. Segments were estimated by the canonical phonemes and segment boundaries by the automatically obtained alignment of the phoneme string to the signal. Speech rate estimates based on all segments and estimates based on vowel segments only are calculated. Duration normalised for inherent phoneme length and for speaker, respectively, is used as well as non-normalised duration. Both duration on a linear scale and on a logarithmic scale are used. All combinations of strategies are included in the annotation, resulting in 16 different speech rate measures for each unit.

Utterance Tier Annotation

The utterance tier annotation includes the variables *speaker sex*, *utterance type* (statement, question/request response, answer/response) and a set of speech rate measures.

Phrase Tier Annotation

The phrase tier annotation includes the variables *phrase type*, *phrase length* (word, syllable and phoneme counts), *prosodic weight* (stress count, focal stress count), and measures of *local speech rate* over each phrase unit and of *pitch dynamism* and *pitch range.*

A pitch extraction algorithm included in the SNACK sound toolkit (Sjölander and Beskow, 2000; Sjölander, 2004) is used to obtain information about the pitch contour of the speech

data. A slope tracking algorithm was used for localising minimum and maximum points or plateaus in the extracted pitch contour. The mean pitch is calculated over each segment of the signal corresponding to a unit over which pitch dynamism and range was to be computed. The sum of the absolute distance between the mean and each extreme value is the *pitch dynamism*. The difference between the largest extreme value and the smallest extreme value is the *pitch range*. In addition to a normal Hz frequency scale, pitch is also measured on the Mel, ERB (equivalent rectangular bandwidth), and semitone scales. The three latter scales are used to give estimates of pitch differences closer to the perceived frequency differences of human listeners.

Word Tier Annotation

In addition to a reference orthographic representation, the variables included in the word tier annotation are *word length* (syllable and phoneme counts), *part of speech*, *morphology* (number, definiteness, case, pronoun form, tense/aspect, mood, voice and degree), *word type* (content word or function word), *word repetitions* (full-form and lexeme), *word predictability* (estimation based on trigram, bigram and unigram statistics from an orthographically transcribed version of the Göteborg Spoken Language Corpus, Allwood et al., 2000), *global word probability* (unigram probability), *the position of the word in the phrase*, *focal stress*, *distance to preceding and succeeding foci* (in number of words), *pause context*, *filled pause context*, *interrupted word context*, *prosodic boundary context* and measures of *local speech rate* over each word unit and of *pitch dynamism* and *pitch range*.

Syllable Tier Annotation

The syllable tier annotation includes the variables *stress*, *accent*, *distance to preceding and succeeding stressed syllable* (in number of syllables), *syllable length* (phoneme count), *syllable nucleus*, *the position of the syllable in the word* and measures of *local speech rate* over each syllable unit.

Phoneme Tier Annotation

On the phoneme level, the annotation includes the *canonical phoneme* and a set of *articulatory features* describing the canonical phoneme, *the position of the phoneme in the syllable* and *in a consonant cluster*, *consonant cluster length* (phoneme count) and the realised *phone*.

Canonical phonological representations of words were collected from a pronunciation lexicon developed at the Centre for Speech Technology (CTT). Phonological forms for words not included in the lexicon were generated using grapheme-to-phoneme rules.

Phonetic transcripts are provided by a system using statistical decoding and a set of correction rules. First, a pronunciation network is created. For each phoneme, a list of possible realisations (tentative phones) is collected from an empirically based realisation list. The phone label set is the same as the phoneme label set and includes 23 vowel symbols and 23 consonant symbols. There is also a place filler **`null`** label in the phone label set used for occupying the phone positions of phonemes with no realisation in the phonetic string.

A finite state transition network is built from the pronunciation net and a set of HMM monophone models (Sjölander and Beskow, 2000; Sjölander, 2003). SNACK tools (Sjölander, 2004) are then used for Viterbi decoding (probability maximisation) given the observation sequence defined by the parameterised speech.

A layer of correction rules are applied to correct some systematic errors made in the Viterbi decoding. The rules use phoneme context (including word stress annotation) and tentative phone context as well as estimated phoneme and tentative phone durations as context.

The correction rules were compiled using a manually transcribed gold standard to detect Viterbi decoder errors and to evaluate the effects of introduced rules. For each phoneme in the canonical representation, the gold standard phone and the phone produced by the decoder were compared. Each type of deviation from the gold standard was investigated with the aim to find consistencies in the context which could be used in formulating correction rules. Rules were written to minimise the phoneme error rate (PER), however with the restriction that the rules should be generally applicable. The final rule system was evaluated on a gold standard different from the development standard used in the development phase. The evaluation showed similar PERs and error distributions for the evaluation gold standard as for the development gold standard, both generally and when separating different speakers. The PER of the autotranscription system when compared to the evaluation gold standard was 14.37%.

Model Performance

The annotation has been used for decision tree model induction (initial results are reported in Jande, 2004). The decision tree pronunciation variation model works with phonemes in a canonical phonemic pronunciation representation as its central units. A vector containing all available context information is connected to each canonical phoneme. For each canonical phoneme, the model makes a decision about the appropriate phone realisation given the context associated with the canonical phoneme.

Decision tree models trained on annotation from elicited monologue showed a PER of 9.91% when evaluated against the same type of data as they were trained on in a tenfold cross validation setting. This meant a 55.25% error reduction compared to using the canonical pronunciation representation for estimating the phonetic realisation.

The decision tree models were pruned to make them more general (less specific to the particular training data from which they were induced). Thus, not all variables were used in the final models. None of the discourse or utterance tier attributes were used in any of the pruned models, probably due to the fact that only one type of speaking style was used. From the phrase, word, syllable and phoneme tiers, many different types of attributes were used. As could be expected, the identity of the canonical phoneme was the primary phone level realisation predictor.

Conclusions

A system for annotation of speech data with variables hypothesised to be important for the pronunciation of words in discourse context has been described. Automatic methods used for obtaining or estimating variables have been presented. The annotation has been used for creating pronunciation variation models in the form of decision trees. The models show a decrease in phone error rate with 55.25% compared to using canonical phonemic word representations from a pronunciation lexicon.

References

Allwood J., Björnberg M., Grönqvist L., Ahlsén E. and Ottesjö C. (2000) The Spoken Language Corpus at the Linguistics Department, Göteborg University. Forum Qualitative Social Research 1.

Aycock J. (1998) Compiling little languages in Python. Proc /th International Python Conference.

Bannert R. and Czigler P. E. (1999) Variations in consonant clusters in standard Swedish. Phonum 7, Umeå University.

Brants T. (2000) TnT – A statistical part-of-speech tagger. Proc ANLP.

Bruce G. (1986) Elliptical phonology. Papers from the Ninth Scandinavian Conference on Linguistics, 86–95.

Finke M. and Waibel A. (1997) Speaking mode dependent pronunciation modeling in large vocabulary conversational speech recognition. Proc Eurospeech, 2379–2382.

Fosler-Lussier E. and Morgan N. (1999). Effects of speaking rate and word frequency on pronunciations in conversational speech. Speech Communication, 29(2–4):137–158.

Gårding E. (1974) Sandhiregler för svenska konsonanter. Svenskans beskrivning 8, 97–106.

Jande P-A (2003a) Evaluating rules for phonological reduction in Swedish. Proc Fonetik, 149–152.

Jande P-A (2003b) Phonological reduction in Swedish. Proc 15th ICPhS, 2557–2560.

Jande, P.-A. (2004). Pronunciation variation modelling using decision tree induction from multiple linguistic parameters. Proc Fonetik, 12–15.

Jurafsky D., Bell A., Gregory M., and Raymond W. (2001). Probabilistic relations between words: Evidence from reduction in lexical production. In Bybee and Hopper (eds) Frequency and the emergence of linguistic structure, 229–254. John Benjamins.

Megyesi B. (2002a) Data-driven syntactic analysis – Methods and applications for Swedish. Ph. D. Thesis. KTH, Stockholm.

Megyesi B. (2002b). Shallow parsing with pos taggers and linguistic features. Journal of Machine Learning Research, 2, 639–668.

Sjölander K. (2003) An HMM-based system for automatic segmentation and alignment of speech. Proc Fonetik, 193–196.

Sjölander K. (2004) The snack sound toolkit. http://www.speech.kth.se/snack/

Sjölander K. and Beskow J. (2000) WaveSurfer - a public domain speech tool. Proc ICSLP, IV, 464–467.

Van Bael C., van den Heuvel H., and Strik H. (2004). Investigating speech style specific pronunciation variation in large spoken language corpora. Proc ICSLP, 586–589.