The patterns of confusions can be modelled by weighted summation of the response probabilities for each vowel in the attended modality [listening (A), lipreading (V)] and a Bayesian auditory-visual integration (AV). For the pooled data, linear regression on this basis gives response probabilities P and determination coefficients r^2 as follows:

$P_{heard} = 0.01 + 0.26\ A + 0.71\ AV\ (r^2 = 0.98)$ and $P_{seen} = -0.00 + 0.57\ V + 0.45\ AV\ (r^2 = 0.94)$.

4 Discussion

As for auditory perception with and without conflicting visual cues and for visual perception alone (lipreading), the patterns of confusion observed here agree closely with those obtained previously (Traunmüller & Öhrström). Now, the novel results obtained in visual perception with conflicting auditory cues demonstrate that a visual percept that may be influenced by audition has to be distinguished from the auditory percept that may be influenced by vision and that the strength of the cross-modal influence is feature-specific in each case.

Based on confusion patterns in consonant perception, it has been claimed that humans behave in accordance with Bayes' theorem (Massaro & Stork, 1998), which allows predicting bimodal response probabilities by multiplicative integration of the unimodal probabilities. Although some of our subjects behaved in agreement with this hypothesis in reporting what they *heard*, the behaviour of most subjects refutes the general validity of this claim, since it shows a substantial additive influence of the auditory sensation. When reporting what they *saw*, all subjects except one showed a substantial additive influence of the visual sensation.

Given the unimodal data included in Tables 1 and 2, Bayesian integration lends prominence to audition in the perception of openness and to vision in roundedness. The data make it clear that an ideal perceiver should rely on audition in the perception of openness, as all subjects did in their auditory judgments, and combine this with the roundedness sensed by vision, since this is more reliable when the speaker's face is clearly visible. Four female and two male subjects behaved in this way to more than 90% in reporting what they *heard* but only one other female subject in reporting what she *saw*.

The results can be understood as reflecting a weighted summation of sensory cues for features such as openness and roundedness, whereby the weight attached reflects the feature-specific reliability of the information received by each sensory modality (cf. Table 3). The between-perceiver variation then reflects differences in the estimation of this reliability.

Acknowledgements

This investigation has been supported by grant 2004-2345 from the Swedish Research Council. I am grateful to Niklas Öhrström for the recordings and for discussion of the text.

References

Massaro, D., 1996. Bimodal speech perception: a progress report. In D.G. Stork & M.E.Hennecke (eds.), *Speechreading by Humans and Machines*. Berlin: Springer, 80-101.

Massaro, D.W. & D.G. Stork, 1998. Speech recognition and sensory integration. *American Scientist 86*, 236-244.

Robert-Ribes, J., M. Piquemal, J-L. Schwartz & P. Escudier, 1996. Exploiting sensor fusion architectures and stimuli complementarity in AV speech recognition. In D.G. Stork & M.E.Hennecke (eds.), *Speechreading by Humans and Machines*. Berlin: Springer, 193-210.

Robert-Ribes, J., J-L. Schwartz, T. Lallouache & P. Escudier, 1998. Complementarity and synergy in bimodal speech: Auditory, visual and audio-visual identification of French oral vowels in noise. *Journal of the Acoustical Society of America 103*, 3677-3689.

Traunmüller, H. & N. Öhrström, in press. Audiovisual perception of openness and lip rounding in front vowels. *Journal of Phonetics*.

Lund University, Centre for Languages & Literature, Dept. of Linguistics & Phonetics 141
Working Papers 52 (2006), 141–144

Knowledge-light Letter-to-Sound Conversion for Swedish with FST and TBL

Marcus Uneson

Dept. of Linguistics and Phonetics, Centre for Languages and Literature, Lund University
marcus.uneson@ling.lu.se

Abstract

This paper describes some exploratory attempts to apply a combination of finite state transducers (FST) and transformation-based learning (TBL, Brill 1992) to the problem of letter-to-sound (LTS) conversion for Swedish. Following Bouma (2000) for Dutch, we employ FST for segmentation of the textual input into groups of letters and a first transcription stage; we feed the output of this step into a TBL system. With this setup, we reach 96.2% correctly transcribed segments with rather restricted means (a small set of hand-crafted rules for the FST stage; a set of 12 templates and a training set of 30kw for the TBL stage).

Observing that quantity is the major error source and that compound morpheme boundaries can be useful for inferring quantity, we exploratively add good precision-low recall compound splitting based on graphotactic constraints. With this simple-minded method, targeting only a subset of the compounds, performance improves to 96.9%.

1 Introduction

A text-to-speech (TTS) system which takes unrestricted text as input will need some strategy for assigning pronunciations to unknown words, typically achieved by a set of letter-to-sound (LTS) rules. Such rules may also help in reducing lexicon size, permitting the deletion of entries whose pronunciation can be correctly predicted from rules alone. Outside the TTS domain, LTS rules may be employed for instance in spelling correction, and automatically induced rules may be interesting for reading research.

Building LTS rules by hand from scratch is easy for some languages (e.g., Finnish, Turkish), but turns out prohibitively laborious in most cases. Data-driven methods include artificial neural networks, decision trees, finite-state methods, hidden Markov models, transformation-based learning and analogy-based reasoning (sometimes in combination). Attempts at fully automatic, data-driven LTS for Swedish include Frid (2003), who reaches 96.9 % correct transcriptions on segment level with a 42000-node decision tree.

2 The present study

The present study tries a knowledge-light approach to LTS conversion, first applied by Bouma (2000) on Dutch, which combines a manually specified segmentation step (by finite-state transducers, FST) and an error-driven machine learning technique (transformation-based learning, TBL). One might think of the first step as redefining the alphabet size, by introducing new, combined letters, and the second as automatic induction of reading rules on that (redefined) alphabet, ordered in sequence of relevance.

For training and evaluation, we used disjoint subsets of a fully morphologically expanded form of Hedelin et al. (1987). The expanded lexicon holds about 770k words (including proper nouns; these and other words containing characters outside the Swedish alphabet in lowercase were discarded).

2.1 Finite-state transduction (FST)

Many NLP tasks can be cast as string transformation problems, often conveniently attacked with context-sensitive rewrite rules (which can be compiled directly into FST). Here, we first use an FST to segment input into segments or letter groups, rather than individual letters. A segment typically corresponds to a single sound (and may have one member only). Treating a sequence of letters as a group is in principle meaningful whenever doing so leads to more predictable behaviour. Clearly, however, there is an upper limit on the number of groups, if the method should justifiably be called 'knowledge-light'. For Swedish, some segments close at hand are {[s,c,h], [s,s], [s,j], [s,h], [c,k], [k], [k,j]...}; the set used in the experiments described here has about 75 members.

Segmentation is performed on a leftmost, longest basis, i.e., that rule is chosen which results in as early a match as possible, the longest possible one if there are several candidates. All following processing now deals with segments rather than individual letters.

After segmentation, markers for begin- and end-of-word are added, and the (currently around 30) hand-written replace rules are applied, again expressed as transducers or compositions of transducers. These context-sensitive replace rules may encode well-known reading rules (in the case of Swedish, for instance '<k> is pronounced /ɕ/ in front of <e,i,y,ä,ö> morpheme-initially'), or try to capture other partial regularities (Olsson 1998). Most rules deal with vowel quantity and/or the <o> grapheme, reflecting typical difficulties in Swedish orthography. The replacement transducer is implemented such that each segment can be transduced at most once. A set (currently around 60) of context-less, catch-all rules provide default mappings. To illustrate the FST steps, consider the word *skärning* 'cut' after each transduction:

```
input:              skärning
segment:            sk-ä-r-n-i-ng
marker:             #-sk-ä-r-n-i-ng-#
transduce:          #-S+<:+r-n-I-N+#
remove marker:      S<:rnIN
```

2.2 Transformation-based learning (TBL)

TBL was first proposed for part-of-speech tagging by Eric Brill (1992). TBL is, generally speaking, a technique for automatic learning of human-readable classification rules. It is especially suited for tasks where the classification of one element depends on properties or features of a small number of other elements in the data, typically the few closest neighbours in a sequence. In contrast to the opaque problem representation in stochastic approaches, such as HMMs, the result of TBL training is a human-readable, ordered list of rules. Application of the rules to new material can again be implemented as FSTs and thus be very fast.

For the present task, we employed the μ-TBL system (Lager 1999). It provides an interface for scripting as well as an interactive environment, and Brill's original algorithm is supplemented by much faster Monte Carlo rule sampling. The templates were taken from Brill (1992), omitting disjunctive contexts (e.g., "A goes to B when C is either 1 or 2 before"), which are less relevant to LTS conversion than to POS tagging.

2.3 Compound segmentation (CS)

The most important error source by far is incorrectly inferred quantity. In contrast to Dutch, for which Bouma reports 99% with the two steps above (and a generally larger setup, with

500 TBL templates), quantity is not explicitly marked in Swedish orthography. One might suspect that this kind of errors might be remedied if compounds and their morpheme boundaries could be identified in a preprocessing step. Many rules are applicable in the beginning or end of morphemes rather than words; we could provide context for more rules if only we knew where the morpheme boundaries are. Compound segmentation (CS) could also help in many difficult cases where the suffix of one component happens to form a letter group when combined with the prefix of the following, as in <matjord>, <polishund>, <bokjägare>. Ideally, segments should not span morpheme boundaries: <sch> should be treated as a segment in <kvällsǀschottis> but not in <kvällsǀchoklad>.

In order to explore this idea while still minimizing dependencies on lexical properties, we implemented a simple compound splitter based on graphotactic constraints. An elaborate variant of such a non-lexicalized method for Swedish was suggested by Brodda (1979). He describes a six-level hierarchy for consonant clusters according to how much information they provide about a possible segmentation point, from certainty (as -rkkl- in <kyrkklocka> 'church bell') to none at all (as -gr- in <vägren> 'verge (road)'). For the purposes of this study, we targeted the safe cases only (on the order of 30-40% of all compounds). Thus, recall is poor but precision good, which at least should be enough to test the hypothesis.

3 Results

3.1 Evaluation measure

The most common LTS evaluation measure is Levenshtein distance between output string and target. For the practical reason of convenient error analysis and comparability with Frid (2003) we follow this, but we note that the measure has severe deficiencies. Thus, all errors are equally important – exchanging [e] for [ə] is considered just as bad as exchanging [t] for [a]. Furthermore, different lexica have different levels of granularity in their transcriptions, leading to rather arbitrary ideas about what 'right' is supposed to mean. For future work, some phonetically motivated distance measure, such as the one suggested by Kondrak (2000), seems a necessary supplement.

Table 1. Results and number of rules for combinations of CS, FST, and TBL. 5-fold cross-validation. Monte Carlo rule sampling. Score threshold (stopping criterion) = 2. The baselines (omitting TBL) are 80.1% (default mappings); 86.6% (FST step only); 88.3% (CS + FST).

Training data		*TBL*		*FST + TBL*		*CS + FST + TBL*	
segments	words	results %	#rules	results %	#rules	results %	#rules
49k	5k	93.8	820	94.9	503	95.5	513
98k	10k	94.1	1131	95.0	761	95.7	809
198k	20k	95.2	1690	95.7	1275	96.5	1250
300k	30k	95.7	2225	96.2	1862	96.9	1756

3.2 Discussion

Some results are given in Table 1. In short, both with and without the TBL steps, adding handwritten rules to the baseline improves system performance (and TBL training time) significantly, as does adding the crude CS algorithm. The number of learnt rules is sometimes high. However, although space constraints do not allow the inclusion of a graph here, rule efficiency declines quickly (as is typical for TBL), and the first few hundred rules are by far the most important. We note that the major error source still is incorrectly inferred quantity.

We have stayed at the segmental level of lexical transcription, with no aim of modelling contextual processes. Although this approach would need (at the very least) postprocessing for many applications, it might be enough for others, such as spelling correction. Result-wise, it seems that the current approach can challenge Frid's (2003) results (96.9% on a much larger (70kw) training corpus), while still retaining the advantage of the more interpretable rule representation. Frid goes on to predict lexical prosody; we hope to get back to this topic.

4 Future directions

Outside incorporating more sophisticated compound splitting, there are several interesting directions. The template set is currently small. Likewise, the feature set for each corpus position may be extended in other ways, for instance by providing classes of graphemes – C and V is a good place to start, but place or manner of articulation for C and frontness for vowels might also be considered. Such classes might help finding generalizing rules over, say, front vowels or nasals, and might help where data is sparse; the extracted rules are also likely to be more linguistically relevant. If so, segments should preferably be chosen such that they fall clear into classes.

Another, orthogonal approach is "multidimensional" TBL (Florian & Ngai 2001), i.e., TBL with more than one variable. For instance, the establishment of stress pattern may determine phoneme transcription, or the other way round. For most TBL systems, rules can change one, prespecified attribute only (although many attributes may provide context). This is true for µ-TBL as well; however, we are currently considering an extension.

Interesting is also the idea to try to predict quantity and stress reductively, with Constraint Grammar-style reduction rules (i.e., "if Y, remove tag X from the set of possible tags"). Each syllable is assigned an initial set of all possible stress levels, a set which is reduced by positive rules ('ending -<ör># has main stress; thus its predecessor does not') as well as negative ('ending -<lig># never takes stress'). µ-TBL conveniently supports reduction rules.

References

Bouma, G., 2000. A finite state and data oriented method for grapheme to phoneme conversion. *Proceedings of the first conference on North American chapter of the Association for Computational Linguistic*, Seattle, WA.

Brill, E., 1992. A simple rule-based part of speech tagger. *Third Conference on Applied Natural Language Processing*, ACL.

Brodda, B., 1979. Något om de svenska ordens fonotax och morfotax: Iakttagelse med utgångspunkt från automatisk morfologisk analys. *PILUS 38*. Institutionen för lingvistik, Stockholms universitet.

Florian, R. & G. Ngai, 2001. Multidimensional Transformation-Based Learning. *Proceedings of the Fifth Workshop on Computational Language Learning (CoNLL-2001)*, Toulouse.

Frid, J., 2003. *Lexical and Acoustic Modelling of Swedish Prosody.* PhD Thesis. Travaux de l'institut de linguistique de Lund 45. Dept. of Linguistics, Lund University.

Hedelin, P., A. Jonsson & P. Lindblad, 1987. *Svenskt uttalslexikon* (3rd ed.). Technical report, Chalmers University of Technology.

Kondrak, G., 2000. A new algorithm for the alignment of phonetic sequences. *Proceedings of the first conference on North American chapter of the ACL*, Morgan Kaufmann Publishers Inc, 288-295.

Lager, T., 1999. The µ-TBL System: Logic Programming Tools for Transformation-Based Learning. *Third International Workshop on Computational Natural Language Learning (CoNLL-1999)*, Bergen.

Olsson, L-J., 1998. Specification of phonemic repesentation, Swedish. *DEL 4.1.3 of EC project "SCARRIE Scandinavian proof-reading tools" (LE3-4239).*

Lund University, Centre for Languages & Literature, Dept. of Linguistics & Phonetics
Working Papers 52 (2006), 145–148

The Articulation of Uvular Consonants: Swedish

Sidney Wood

Dept. of Linguistics and Phonetics, Centre for Languages and Literature, Lund University
sidney.wood@ling.lu.se

Abstract

The articulation of uvular consonants is studied with particular reference to quantal aspects of speech production. Data from X-ray motion films are presented. Two speakers of Southern Swedish give examples of [ʀ]. The traditional view, that uvular consonants are produced by articulating the tongue dorsum towards the uvula, is questioned, and theoretical considerations point instead to the same upper pharyngeal place of articulation as for [o]-like vowels. The X-ray films disclose that these subjects did indeed constrict the upper pharynx for [ʀ].

1 Introduction

1.1 The theory of uvular articulations

This study begins by questioning the classical account of uvular consonant production (e.g. Jones, 1964), that the tongue dorsum is raised towards the uvula, and that the uvula vibrates for a rolled [ʀ]. Firstly, it is not clear how a vibrating uvula would produce the acoustic energy of a typical rolled [ʀ]. A likely process exploits a Bernoulli force in the constricted passage to chop the voiced sound into pulses when air pressure and tissue elasticity are suitably balanced, which requires that intermittent occlusion is possible between pulses. Unfortunately, there are free air passages either side of the uvula that should prevent this from happening. Secondly, these same passages should likewise prevent complete occlusion for a uvular stop, and they should also prevent a Reynolds number becoming sufficiently small for the turbulence of uvular fricatives.

If the uvula is not a good place for producing consonants known as "uvular", how else might they be produced? Wood (1974) observed that the spectra of vowel-to-consonant transitions immediately adjacent to uvular consonants were very similar to the spectra of [o ɔ]-like vowels, or to their respective counterparts [ɤ ʌ], and concluded that they shared the same place of location, i.e. the upper pharynx, confirmed for [o ɔ]-like vowels by Wood (1979). Mrayati et al. (1988) studied the spectral consequences of systematic deformations along an acoustic tube, and also concluded that the upper pharynx was a suitable location for these same consonants and vowels. Observations like this are obviously relevant for discussions of the quantal nature of speech (Stevens 1972, 1989). Clarifying the production of uvular consonants is not just a matter of correcting a possible misconception about a place of articulation. It concerns fundamental issues of phonetic theory.

1.2 This investigation

The uvular articulations were analysed from cinefluorographic films, a method that enables simultaneous articulatory activity to be observed in the entire vocal tract, and is therefore suitable for studying the tongue manoeuvres associated with uvular consonants. Two undisputed sources of uvular consonants are [ʀ] in southern Swedish, and [ʀ q χ] in West Greenlandic Inuit. The subjects of the films are native speakers of these languages. Examples