Modeling the language assessment process and result: Proposed architecture for automatic oral proficiency assessment

Gina-Anne Levow and Mari Broman Olsen
University of Maryland Institute for Advanced Computer Studies
College Park, MD 20742
{gina,molsen}@umiacs.umd.edu

Abstract

We outline challenges for modeling human language assessment in automatic systems, both in terms of the process and the reliability of the result. We propose an architecture for a system to evaluate learners of Spanish via the Computerized Oral Proficiency Instrument, to determine whether they have 'reached' or 'not reached' the Intermediate Low level of proficiency, according to the American Council on the Teaching of Foreign Languages (ACTFL) Speaking Proficiency Guidelines. Our system divides the acoustic and non-acoustic features, incorporating human process modeling where permitted by the technology and required by the domain. We suggest machine learning techniques applied to this type of system permit insight into yet unarticulated aspects of the human rating process.

1 Introduction

Computer-mediated language assessment appeals to educators and language evaluators because it has the potential for making language assessment widely available with minimal human effort and limited expense. Fairly robust results ($\kappa \approx 0.8$) have been achieved in the commercial domain modeling the human rater results, with both the Electronic Essay Rater (e-rater) system for written essay scoring (Burstein et al., 1998), and the PhonePass pronunciation assessment (Ordinate, 1998).

There are at least three reasons why it is not possible to model the human rating process. First, there is a mismatch between what the technology is able to handle and what people manipulate, especially in the assessment of speech features. Second, we lack a well-articulated model of the human process, often characterized as holistic. Certain assessment features have been identified, but their relative importance is not clear. Furthermore, unlike automatic assessments, human raters of oral proficiency exams are trained to focus on competencies, which are difficult to enumerate. In contrast, automatic assessments of spoken language fluency typically use some type of error counting, comparing duration, silence, speaking rate and pronunciation mismatches with native speaker models.

There is, therefore, a basic tension within the field of computer-mediated language assessment, between modeling the assessment process of human raters or achieving comparable, consistent assessments, perhaps through different means. Neither extreme is entirely satisfactory. A spoken assessment system that achieves human-comparable performance based only, for example, on the proportion of silence in an utterance would seem not to be capturing a number of critical elements of language competence, regardless of how accurate the assessments are. Such a system would also be severely limited in its ability to provide constructive feedback to language learners or teachers. The e-rater system has received similar criticism for basing essay assessments on a number of largely lexical features, rather than on a deeper, more human-style rating process.

Thirdly, however, even if we could articulate and model human performance, it is not clear that we want to model all aspects of the human rating process. For example, human performance varies due to fatigue. Transcribers often inadvertently correct examinees' errors of omitted or incorrect articles, conjugations, or affixes. These mistakes are a natural effect of a cooperative listener; however, they result in an over-optimistic assessment of the speaker's actual proficiency. We arguably do not wish to build this sort of cooperation into an automated assessment system, though it is likely desirable for other sorts of human-computer interaction systems.

Furthermore, if we focus on modeling human processes we may end up underutilizing the technology. Balancing human-derived features with machine learning techniques may actually allow us to discuss more about the human rating process by making the entire process available for inspection and evaluation. For example, if we are able to articulate human rating features, machine learning techniques may allow us to 'learn' the relative weighting of these features for a particular assessment value.

2 Modeling the rater

2.1 Inference & Inductive Bias

Research in machine learning has demonstrated the need for some form of inductive bias, to limit the space of possible hypotheses the learning system can infer. In simple example-based concept learning, concepts are often restricted to certain classes of Boolean combinations, such as conjuncts of disjuncts, in order to make learning tractable. Recent research in automatic induction of context-free grammars, a topic of more direct interest to language learning and assessment, also attests to the importance of structuring the class of grammars that can be induced from a data set. For instance Pereira and Schabes (1992) demonstrate that a grammar learning algorithm with a simple constraint on binary branching (CNF) achieves less than 40% accuracy after training on an unbracketed corpus.

Two alternatives achieve comparable increases in grammatical accuracy. Training on partially bracketed corpora - providing more supervision and a restriction on allowable grammars - improves to better than 90%. (DeMarcken, 1995) finds that requiring binary branching, as well as headedness and head projection restrictions on the acquirable grammar, leads to similar improvements. These results argue strongly that simply presenting raw text or feature sequences to a machine learning program to build an automatic rating system for language assessment is of limited utility. Results will be poorer and require substantially more training data than if some knowledge of the task or classifier end-state based in human knowledge and linguistic theory is applied to guide the search for classifiers.

2.2 Encoding Linguistic Knowledge

Why, then, if it is necessary to encode human knowledge in order to make machine learning practical, do we not simply encode each piece of the relevant assessment knowledge from the person to the machine? Here again parallels with other areas of Natural Language Processing (NLP) and Artificial Intelligence (AI) provide guidance. While both rule-based, hand-crafted grammars and expert systems have played a useful role, they require substantial labor to construct and become progressively more difficult to maintain as the number of rules and rule interactions increases. Furthermore, this labor is not transferable to a new (sub-)language or topic and is difficult to encode in a way that allows for graceful degradation.

Another challenge for primarily hand-crafted approaches is identifying relevant features and their relative importance. As is often noted, human assessment of language proficiency is largely holistic. Even skilled raters have difficulty identifying and quantifying those features used and their weights in determining an assessment. Finally, even when identifiable, these features may not be directly available to a computer system. For instance, in phonology, human listeners perceive categorical distinctions between phonemes (Eimas et al., 1971; Thibodeau and Sussman, 1979) whereas acoustic measures vary continuously.

We appeal to machine learning techniques in the acoustic module, as well as in the pooling of information from both acoustic and non-acoustic features.

3 Domain: The Computerized Oral Proficiency Instrument

The Center for Applied Linguistics in Washington, D.C. (CAL) has developed or assisted in developing simulated oral proficiency interview (SOPI) tests for a variety of languages, recently adapting them to a computer-administered format, the COPI. Scoring at present is done entirely by human raters. The Spanish version of the COPI is in the beta-test phase; Chinese and Arabic versions are under development. All focus on assessing proficiency at the Intermediate Low level, defined by the American Council on the Teaching of Foreign Languages (ACTFL)

Speaking Proficiency Guidelines (ACT, 1986), a common standard for passing at many high schools. We focus on Spanish, since we will have access to real data. Our goal is to develop a system with a high interannotator agreement with human raters, such that it can replace one of the two or three raters required for oral proficiency interview scoring.

With respect to non-acoustic features, our domain is tractable, for current natural language processing techniques, since the input is expected to be (at best) sentences, perhaps only phrases and words at Intermediate Low. Although the tasks and the language at this level are relatively simple, the domain varies enough to be interesting from a research standpoint: enumerating items in a picture, leaving a answering machine message, requesting a car rental, giving a sequence of directions, and describing one's family, among others. These tasks elicit a more varied, though still topically constrained, vocabulary. They also allow the assessment of the speaker's grasp of target language syntax, and, in the more advanced tasks, discourse structure and transitions. The COPI, therefore, provides a natural domain for rating non-native speech on both acoustic and non-acoustic features. These subsystems differ in terms of how amenable they are to machine modeling of the human process, as outlined below.

4 Acoustic Features: The Speech Recognition Process

In the last two decades significant advances have been made in the field of automatic speech recognition (SR), both in commercial and research domains. Recently, research interest in recognizing non-native speech has increased, providing direct comparisons of recognition accuracy for non-native speakers at different levels of proficiency (Byrne et al., 1998). Tomokiyo (p.c.), in experiments with the JANUS (Waibel et al., 1992) speech system developed by Carnegie Mellon University, reports that systems with recognition accuracies of 85% for native speech perform at 40-50% for high fluency L2 learners (German, Tomokiyo, p.c.) and 30% for medium fluency speech (Japanese, Tomokiyo, p.c.).

However, the current speech recognition technology makes little or no effort to model the human auditory or speech understanding process. Furthermore, standard SR approaches to speaker adaptation rely on relatively large amounts (20-30 minutes) of fixed, recorded speech (Jecker, 1998) to modify the underlying model, say in the case of accented speech, again unlike human listeners.

While a complete reengineering of speech recognition is beyond the scope of our current project, we do attempt to model the human assessor's approach to understanding non-native speech. The SR system allows us two points of access through which linguistic knowledge of L2 phonology and grammar can be applied to improve recognition: the lexicon and the speech recognizer grammar.

4.1 Lexicon: Transfer-model-based Phonological Adaptation

Since we have too little data for conventional speaker adaptation (less than 5 minutes of speech per examinee), we require a principled way of adapting an L1 or L2 recognizer model to non-native learner's speech that places less reliance upon recorded training data. We know that the pronunciation reflects native language influence, most notably at early stages (Novice and Intermediate), with which we are primarily concerned. Following the L2 transfer acquisition model, we assume that the L2 speaker, in attempting to produce target language utterances, will be influenced by L1 phonology and phonotactics. Thus, rather than being random divergences from TL pronunciation, errors should be closer to L1 phonetic realizations.

To model these constraints, we will employ two distinct speech recognizers that can be employed to recognize L2 speech, produced by adaptations specific to Target Language (TL) and Source Language (SL). We propose to use language identification technology to arbitrate between the two sets of recognizer results, based on a sample of speech, either counting to 20, or a short read text. Since we need to choose between an underlying TL phonological model and one based on the SL, we will make the selection based on the language identification decision as to the apparent phonological identity of the sample as SL or TL, based on the sample's phonological and acoustic features (Berkling and Barnard, 1994; Hazen and

Zue, 1994; Kadambe and Hieronymus, 1994; Muthusamy, 1993). Parameterizing phonetic expectation based on a short sample of speech (Ladefoged and Broadbent, 1957) or expectations in context (Ohala and Feder, 1994) mirrors what people do in speech processing generally, independent of the rating context.

4.2 An acoustic grammar: Modeling the process

Modeling the grammar of a Novice or Intermediate level L2 speaker for use by a speech recognizer is a challenging task. As noted in the ACTFL guidelines, these speakers are frequently inaccurate. However, to use the content of the speech in the assessment, we need to model human raters, who recognize even errorful speech as accurately and completely as possible. Speech recognizers work most effectively when perplexity is low, as is the case when the grammar and vocabulary are highly constrained. However, speech recognizers also recognize what they are told to expect, often accepting and misrecognizing utterances when presented with out-of-vocabulary or out-of-grammar input. We must balance these conflicting demands.

We will take advantage of the fact that this task is being performed off-line and thus can tolerate recognizer speeds several times real-time. We propose a multi-pass recognition process with step-wise relaxation of grammatical constraints. The relaxed grammar specifies a noun phrase with determiner and optional adjective phrase but relaxes the target language restrictions on gender and number agreement among determiner, noun, and adjective and on position of adjective. Similar relaxations can be applied to other major constructions, such as verbs and verbal conjugations, to pass, without "correcting", utterances with small target language inaccuracies. For those who would not reach such a level, and for tasks in which sentence-level structure is not expected, we must relax the grammar still further, relying on rejection at the first pass grammar to choose grammars appropriately. Successive relaxation of the grammar model will allow us to balance the need to reduce perplexity as much as possible with the need to avoid over-predicting and thereby correcting the learner's speech.

4.3 Acoustic features: Modeling the result

Research in the area of pronunciation scoring (Rypa, 1996; Franco et al., 1997; Ehsani and Knodt, 1998) has developed both direct and indirect measures of speech quality and pronunciation accuracy, none of which seem to model human raters at any level. The direct measures include calculations of phoneme error rate, computed as divergence from native speaker model standards, and number of incorrectly pronounced phonemes. The indirect measures attempt to capture some notion of fluency and include speaking rate, number and length of pauses or silences, and total utterance length. Analogous measures should prove useful in the current assessment of spoken proficiency. In addition, one could include, as a baseline, a human-scored measure of perceived accent or lack of fluency. A final measure of acoustic quality could be taken from the language identification process used in the arbitration phase, as to whether the utterance was more characteristic of the source or target language. In our samples of Intermediate Low passing speech we identify, for example, large proportions of silence to speech both between and within sentences. Some sentences are more than 50% silence.

5 Natural Language Understanding: Linguistic Features Assessment

In the non-acoustic features, we have a fairly explicit notion of generative competence and a reasonable way of encoding syntax in terms of Context-Free Grammars (CFGs) and semantics via Lexical Conceptual Structures (LCSs). We do not know, however, the relative importance of different aspects of this competence in determining reached/not reached for particular levels in an assessment task. Therefore, we apply machine learning techniques to pool the human-identified features, generating a machine-based model of process which is fully explicit and amenable to evaluation.

The e-rater system, deployed by the Educational Testing Service (ETS) incorporates more than 60 variables based on properties used by human raters and divided into syntactic, rhetorical and topical content categories. Although the features deal with suprasentential structure, the reported variables (Burstein et al., 1998) are identified via lexical information and shallow constituent parsing, arguably not modeling the human process.

We attempt to model the features based on a deeper analysis of the structure of the text at various levels. We propose to parallel the architecture of the Military Language Tutoring system (MILT), developed jointly by the University of Maryland and Micro Analysis and Design corporation under army sponsorship. MILT provides a robust model of errors from English speakers learning Spanish and Arabic, identifying lexical and syntactic characteristics of short texts, as well as low-level semantic features, a prerequisite for more sophisticated inferencing (Dorr et al., 1995; Weinberg et al., 1995). At a minimum, the system will provide linguistically principled feedback on errors of various types, rather than providing system error messages, or crashing on imperfect input.

Our work with MILT and the COPI beta-test data suggests that relevant features may be found in each of four main areas of spoken language processing: acoustic, lexical, syntactic/semantic, and discourse. In order to automate the assessment stage of the oral proficiency exam, we must identify features of the L2 examinees' utterances that are correlated with different ratings and that can be extracted automatically. If we divide language up into separate components, we can describe a wide range of variation within a bounded set of parameters within these components. We can therefore build a cross-linguistically valid meta-interpreter with the properties we desire (compactness, robustness and extensibility). This makes both engineering and linguistic sense. Our system treats the constraints as submodules, able to be turned on or off, at the instructor's choice, made based on, e.g., what is learned early, and the level of correction desired. The MILT-style architecture allows us to make use of the University of Maryland's other parsing and lexicon resources, including large scale lexica in Spanish and English.

5.1 Lexical features

One would expect command of vocabulary to be a natural component of a language learner's proficiency in the target language. A variety of automatically extractable measures provide candidate features for assessing the examinee's lexical proficiency. In addition, the structure of the tasks in the examination allows for testing of extent of lexical knowledge in restricted common topics. For instance, the student may be asked to count to twenty in the target language or to enumerate the items in a pictured context, such as a classroom scene. Within these tasks one can test for the presence and number of specific desired vocabulary items, yielding another measure of lexical knowledge.

Simple measures with numerical values include number of words in the speech sample and number of distinct words. In addition, examinees at this level frequently rely on vocabulary items from English in their answers.

A deeper type of knowledge may be captured by the lexicon in Lexical Conceptual Structure (LCS) (Dorr, 1993b; Dorr, 1993a; Jackendoff, 1983). The LCS is an interlingual framework for representing semantic elements that have syntactic reflexes.[1] LCSs have been ported from English into a variety of languages, including Spanish, requiring a minimum of adaptation in even unrelated languages (e.g. Chinese (Olsen et al., 1998)). The representation indicates the argument-taking properties of verbs (*hit* requires an object; *smile* does not), selectional constraints (the subject of *fear* and the object of *frighten* are animate), thematic information of arguments (the subject of *frighten* is an agent; the object is a patient) and classification information of verbs (motion verbs like *go* are conceptually distinct from psychological verbs like *fear/frighten*; *run* is a more specific type of motion verb than *go*). Each information type is modularly represented and therefore may be separately analyzed and scored.

5.2 Syntactic features

We adopt a generative approach to grammar (Government and Binding, Principles and Parameter, Minimalism) principles. In these models, differences in the surface structure of languages can be reduced to a small number of modules and parameters. For example, although Spanish and English both have subject-verb-object (SVO) word order, the relative ordering of many nouns and adjectives differs (the

[1] That is, the LCS does not represent non-syntactic aspects of 'meaning', including metaphor and pragmatics.

"head parameter").[2] In English the adjective precedes the noun, whereas Spanish adjectives of nationality, color and shape regularly follow nouns (Whitley, 1986) [pp. 241-2]. The MILT architecture allows us both to enumerate errors of these types, and parse data that includes such errors. We will also consider measures of number and form of distinct construction types, both attempted and correctly completed. Such constructs could include simple declaratives (subject, verb, and one argument), noun phrases with both determiner and adjective, with correct agreement and word order, questions, and multi-clause sentences.

5.3 Semantic features

Like the syntactic information, lexical information can be used modularly to assess a variety of properties of examinees' speech. The Lexical Conceptual Structure (LCS) allows principles of robustness, flexibility and modularity to apply to the semantic component of the proposed system. The LCS serves as the basis of several different applications, including machine translation and information retrieval as well as foreign language tutoring. The LCS is considered a subset of mental representation, that is, the language of mental representation as realized in language (Dorr et al., 1995). Event types such as event and state, are represented in primitives such as GO, STAY, BE, GO-EXT and ORIENT, used in spatial and other 'fields'. As such, it allows potential modeling of human rater processes.

The LCS allows various syntactic forms to have the same semantic representation, e.g. *Walk to the table and pick up the book, Go to the table and remove the book*, or *Retrieve the book from the table.* COPI examinees are also expected to express similar information in different ways. We propose to use the LCS structure to handle and potentially enumerate competence in this type of variation.

Stored LCS representations may also handle hierarchical relations among verbs, and divergences in the expression of elements of meaning, sometimes reflecting native language word order. The modularity of the system allows us to tease the semantic and syntactic features apart, giving credit for the semantic expression, but identifying divergences from the target L2.

5.4 Discourse features

Since the ability to productively combine words in phrase and sentence structures separates the Intermediate learner from the Novice, features that capture this capability should prove useful in semi-automatic assessment of Intermediate Low level proficiency, our target level. According to the ACTFL Speaking Proficiency Guidelines, Intermediate Low examinees begin to compose sentences; full discourses do not emerge until later levels of competence. Nevertheless, we want both to give credit for any discourse-level features that surface, as well as to provide a slot for such features, to allow scalability to more complex tasks and higher levels of competence. We will therefore develop discourse and dialog models, with the appropriate and measurable characteristics. Many of these can be lexically or syntactically identified, as the ETS GMAT research shows (Burstein et al., 1998). Our data might include uses of discourse connectives (*entonces* 'then; in that case' *pero* 'but'; *es que* 'the fact is'; *cuando* 'when'), other subordinating structures (*Yo creo que* 'I think that') and use of pronouns instead of repeated nouns. The discourse measures can easily be expanded to cover additional, more advanced constructions that are lexically signaled, such as the use of subordination or paragraph-level structures.

6 Machine learning

While the above features capture some measure of target language speaking proficiency, it is difficult to determine *a priori* which features or groups of features will be most useful in making an accurate assessment. In this work, human assessor ratings for those trained on the Speaking Proficiency Guidelines will be used as the "Gold Standard" for determining accuracy of automatic assessment. We plan to apply machine learning techniques to determine the relative importance of different feature values in rating a speech sample.

The assessment phase goes beyond the current work in test scoring, combining recognition of acoustic features, such as the Automatic Spoken Language Assessment by Telephone (ASLAT) or PhonePass (Ordinate, 1998)

[2] Other parameters deal with case, theta-role assignment, binding, and bounding.

with aspects of the syntactic, discourse, and semantic factors, as in e-rater. Our goal is to have the automatic scoring system mirror the outcome of raters trained in the ACTFL Guidelines (ACT, 1986), to determine whether examinees did or did not reach the Intermediate Low level. We also aim to make the process of feature weighting transparent, so that we can determine whether the system provides an adequate model of the human rating process. We will evaluate quantitatively the extent to which machine classification agrees with human raters on both acoustic and non-acoustic properties alone and separately. We will also evaluate the process qualitatively with human raters.

We plan to exploit the natural structuring of the data features through decision trees or a small hierarchical "mixture-of-experts"- type model (Quinlan, 1993; Jacobs et al., 1991; Jordan and Jacobs, 1992). Intuitively, the latter approach creates experts (machine-learning trained classifiers) for each group of features (acoustic, lexical, and so on). The correct way of combining these experts is then acquired though similar machine learning techniques. The organization of the classifier allows the machine learning technique at each stage of the hierarchy to consider fewer features, and thus, due to the branching structure of tree classifier, dramatically fewer classifier configurations.

Decision tree type classifiers have an additional advantage: unlike neural network or nearest neighbor classifiers, they are easily interpretable by humans. The trees can be rewritten trivially as sequences of if-then rules leading to a certain classification. For instance, in the assessment task, one might hypothesize a rule of the form: IF silence > 20% of utterance, THEN Intermediate Low NOT REACHED. It is thus possible to have human raters analyze how well the rules agree with their own intuitions about scoring and to determine which automatic features play the most important role in assessment.

7 Conclusions

We have outlined challenges for modeling the human rating task, both in terms of process and result. In the domain of acoustic features and speech recognition, we suggest the technology currently does not permit complete modeling of the rating process. Nevertheless, the paucity of data in our domain requires us to adopt the transfer model of speech, which permits automatic adaptation to errorful speech. In addition, our recognizer incorporates a relaxed grammar, permitting input to vary from the target language at the lexical and syntactic levels. These adaptations allow us to model human perception and processing of (non-native) speech, as required by our task. In the non-acoustic domain, we also adopt machine learning techniques to pool the relevant human-identified features. As a result, we can learn more about the feature-weighting in the process of tuning to an appropriate level of reliability with the results of human raters.

References

1986. Proficiency guidelines. American Council for the Teaching of Foreign Languages.

Kay M. Berkling and Etienne Barnard. 1994. Language identification of six languages based on a common set of broad phonemes. In *Proceedings of ICSLP [ICS94]*, pages 1891–1894.

Jill Burstein, Karen Kukich, Susanne Wolff, Chi Lu, Martin Chodorow, Lisa Braden-Harder, and Mary Dee Harris. 1998. Automated scoring using a hybrid feature identification technique. In *ACL/COLING 98*, pages 206–210, Montreal, Canada, August 10-14.

William Byrne, Eva Knodt, Sanjeev Khudanpur, and Jared Bernstein. 1998. Is automatic speech recognition ready for non-native speech? a data collection effort and initial experiments in modeling conversational hispanic english. In *Proceedings of STiLL (Speech Technology in Language Learning)*, Marholmen, Sweden. European Speech Communication Association, May.

Carl DeMarcken. 1995. Lexical heads, phrase structure, and the induction of grammar. In *Proceedings of Third Workshop on Very Large Corpora*, Cambridge, MA.

Bonnie J. Dorr, Jim Hendler, Scott Blanksteen, and Barrie Migdaloff. 1995. Use of LCS and Discourse for Intelligent Tutoring: On Beyond Syntax. In Melissa Holland, Jonathan Kaplan, and Michelle Sams, editors, *Intelligent Language Tutors: Balancing Theory and*

Technology, pages 289–309. Lawrence Erlbaum Associates, Hillsdale, NJ.

Bonnie J. Dorr. 1993a. Interlingual Machine Translation: a Parameterized Approach. *Artificial Intelligence*, 63(1&2):429–492.

Bonnie J. Dorr. 1993b. *Machine Translation: A View from the Lexicon.* The MIT Press, Cambridge, MA.

Farzad Ehsani and Eva Knodt. 1998. Speech technology in computer-aided language learning: Strengths and limitations of a new CALL paradigm. *Language Learning & Technology*, 2(1), July.

Peter D. Eimas, Einar R. Siqueland, Peter Jusczyk, and James Vigorito. 1971. Speech perception in infants. *Science*, 171(3968):303–306, January.

H. Franco, L. Neumeyer, Y. Kim, and O. Ronen. 1997. Automatic pronunciation scoring for language instruction. In *Proceedings of ICASSP*, pages 1471–1474, April.

Timothy J. Hazen and Victor W. Zue. 1994. Recent improvements in an approach to segment-based automatic language identification. In *Proceedings of ICSLP [ICS94]*, pages 1883–1886.

Ray Jackendoff. 1983. *Semantics and Cognition.* The MIT Press, Cambridge, MA.

R. A. Jacobs, M. I. Jordan, S. J. Nowlan, and G. E. Hinton. 1991. Adaptive mixtures of local experts. *Neural Computation*, 3(1):79–87.

D. Jecker. 1998. Speech recognition – performance tests. *PC Magazine*, 17, March.

M. I. Jordan and R. A. Jacobs. 1992. Hierarchies of adaptive experts. In *Nips4.*

Shubha Kadambe and James L. Hieronymus. 1994. Spontaneous speech language identification with a knowledge of linguistics. In *Proceedings of ICSLP [ICS94]*, pages 1879–1882.

P. Ladefoged and D.E. Broadbent. 1957. Information conveyed by vowels. *Journal of the Acoustical Society of America*, 29:98–104.

Yeshwant K. Muthusamy. 1993. *A Segmental Approach to Automatic Language Identification.* Ph.D. thesis, Oregon Graduate Institute of Science & Technology, P.O. Box 91000, Portland, OR 97291-1000.

J.J. Ohala and D. Feder. 1994. Listeners' normalization of vowel quality is influenced by restored consonantal context. *Phonetica*, 51:111–118.

Mari Broman Olsen, Bonnie J. Dorr, and Scott C. Thomas. 1998. Enhancing Automatic Acquisition of Thematic Structure in a Large-Scale Lexicon for Mandarin Chinese. In *Proceedings of the Third Conference of the Association for Machine Translation in the Americas, AMTA-98, in Lecture Notes in Artificial Intelligence, 1529*, pages 41–50, Langhorne, PA, October 28-31.

Ordinate. 1998. The PhonePass Test. Technical report, Ordinate Corporation, Menlo Park, CA, January.

Fernando Pereira and Yves Schabes. 1992. Inside-outside reestimation from partially bracket corpora. In *Proceedings of the 30th Annual Meeting of the Association for Computational Linguistics*, pages 128–135, Newark, DE.

J. R. Quinlan. 1993. *C4. 5: Programs for Machine Learning.* Morgan Kaufmann, San Mateo, CA.

M. Rypa. 1996. VILTS: The voice interactive language training system. In *Proceedings of CALICO*, July.

Linda M. Thibodeau and Harvey M. Sussman. 1979. Performance on a test of categorical perception of speech in normal and communication disordered children. *Phonetics*, 7(4):375–391, October.

A. Waibel, A.N. Jain, A. McNair, J. Tebelsis, L. Osterholtz, H. Saito, O. Schmidbauer, T. Sloboda, and M. Woszczyna. 1992. JANUS: Speech-to-Speech Translation Using Connectionist and Non-Connectionist Techniques. In J.E. Moody, S.J. Hanson, and R.P. Lippman, editors, *Advances in Neural Information Processing Systems 4.* Morgan Kaufmann.

Amy Weinberg, Joseph Garman, Jeffery Martin, and Paola Merlo. 1995. Principle-Based Parser for Foreign Language Training in German and Arabic. In Melissa Holland, Jonathan Kaplan, and Michelle Sams, editors, *Intelligent Language Tutors: Theory Shaping Technology*, pages 23–44. Lawrence Erlbaum Associates, Hillsdale, NJ.

Stanley M. Whitley. 1986. *Spanish/English Contrasts: A Course in Spanish Linguistics.* Georgetown University Press, Washington, D.C.