Strategies for Automatic Multi-Tier Annotation of Spoken Language Corpora

Steven Greenberg

The Speech Institute
Oakland, CA 94619 USA
steveng@cogsci.berkeley.edu

Abstract

Spoken corpora of the future will be annotated at multiple levels of linguistic organization largely through automatic methods using a combination of sophisticated signal processing, statistical classifiers and expert knowledge. It is important that annotation tools be adaptable to a wide range of languages and speaking styles, as well as readily accessible to the speech research and technology communities around the world. This latter objective is of particular importance for minority languages, which are less likely to foster development of sophisticated speech technology without such universal access.

1. The Vision

Sometime in the (hopefully, not-too-distant) future there will exist annotated corpora of *spoken* material for most of the principal and leading minority languages of the world. Such corpora will be used by speech technologists and scientists alike, and will come to embody much of what is known about a particular language in terms of its linguistic properties, particularly at the phonetic, prosodic and lexical levels. Speech technologists will use such materials for training automatic speech recognition systems and developing realistic-sounding synthesis. Foreign-language instructors will use such corpora for improving the fluency and pronunciation of their students, while speech pathologists will help patients improve their articulation using such materials. The hearing impaired will benefit as well, as auditory prostheses will be tuned for specific languages and listening environments. Linguistics and speech science pedagogy will change dramatically (and for the better).

This is a vision for the future. What will it take to insure this vision comes to light? And what is being done currently to accomplish such objectives?

2. Spoken Corpora – Past and Present

Linguistic corpora have traditionally been annotated almost exclusively in words. When the Linguistic Data Consortium (LDC) [33], European Language Resources Association (ELRA) [15] or comparable institution releases a corpus, the annotation generally contains a word transcript, along with conversation break points (a.k.a. "turns") and some information about the speakers involved (e.g., age, gender, dialect). Unless the material consists exclusively of read text (e.g., the Wall Street Journal corpus) a word-level transcription has to be prepared, usually by highly trained (but poorly paid) individuals who toil hour after hour deciphering words contained in the recordings. In addition to words, many contemporary corpora also include non-speech sounds such as laughing, coughing and hissing, as well as signals of non-human origin, such as door slams, fan noise and the like. A recent trend has been to label speech in terms of its pragmatic properties regardless of whether words are involved or not (e.g., [16][35]). Hesitations like "uhmmm," or acknowledgements such as "uh huh" are annotated under the rubric of "dysfluencies" or "filled pauses." Because such non-words form an integral part of an utterance they often provide crucial information for reconstructing the speech act for speech scientists and technologists alike.

It is a non-trivial exercise to collect and annotate a corpus, even when the transcription is restricted to the word level. Many months of intensive labor are involved. In addition to the data collection itself, the recordings must be previewed and sorted into appropriate categories. For any large-scale corpus involving many different speakers the effort required often takes years to complete.

One of the earliest efforts of this kind was a corpus of spontaneous telephone dialogues (Am. English) collected at Bell Laboratories in the late 1920's [17]. What is remarkable about this corpus is that it contains not just words, but also their phonetic constituents. Performed before the days of the tape recorder and computer, Bell scientists relied on cylinder recordings and trained transcribers – truly a Herculean task given the technology then available. Not until the mid-1990's was a comparable effort made for a broader range of spontaneous American English telephone dialogues using the SWITCHBOARD corpus [18]. Five hours of material was phonetically labeled and segmented at the International Computer Science Institute (ICSI) [19], and an additional hour annotated with respect to prosodic (syllable) prominence [22]. The Oregon Graduate Institute's Center for Spoken Language mounted a smaller-scale phonetic transcription effort (in English) that also includes brief (60-second) telephone monologues for non-English languages as well (OGI-TS corpus) [31].

For British English, two spoken corpora are of particular note. The London-Lund Corpus of Spoken English (begun at University College London in the late 1950's and ultimately expanded to include Lund University in the mid 1970's) comprises ca. 1,500,000 words. Its focus is on words, prosody, paralinguistics and grammar rather than on phonetics [41]. The second collection effort of interest is the British National Corpus (BNC) [2]. Although most of the BNC is in written form, a subset (ca. 10% or 10,000,000 words) is spoken and annotated automatically in terms of part of speech (POS), but not phonetically or prosodically.

Apart from English, large-scale corpora existed until recently only for German. Principal among these are the Bavarian Archive for Speech Signals (BAS) at Ludwig Maximilians Universität (Munich) [37] and the Kiel Corpus (comprising the Kiel Corpus of Read Speech and the Kiel Corpus of Spontaneous Speech) from Kiel University [25]. Both of these corpora are annotated at the word and phonetic-segment level.

Over the past five years major efforts have been launched for languages other than English and German. The Spoken Dutch Corpus (SDC) contains ca. 1,000 hours of Dutch (Netherlands) and Flemish (Belgium) [7]. Some of the material is annotated at the broad phonetic and prosodic (syllable prominence) level [6][7][12]. A separate effort has been launched for Swedish [1], including the use of video to capture the visual component of speaking. The other principal effort has been made for Japanese – The Corpus of Spontaneous Japanese (CSJ) comprises about 700 hours of material, mostly monologues of lecture presentations and news commentaries [35]. About 45 hours of this material has been quasi-manually labeled at the phonemic and prosodic (ToBI) levels [26]. The remainder has been labeled at the lexical and morphological levels, with automatic POS tagging. A portion of the phonetic labeling was performed manually, with the remainder automatically generated using Viterbi alignment methods [26].

Although the current spoken-corpus projects will undoubtedly provide invaluable material for speech technology and science, they do not, in and of themselves, address the general issue of widespread (and affordable) annotation for the world's languages. That relatively few countries have developed spoken language corpora testifies to the intensive effort and cost such projects entail. Some other strategy is required, one that provides a cost-effective, efficient means with which to annotate spoken language.

3. Phonetic Segment Annotation

Word-level annotation remains the primary objective of spoken corpus development. Traditionally, lexical transcription has been performed entirely by hand, but it is likely that at least some annotation will be performed by automatic methods in the not-too-distant future (albeit with manual verification). One means by which to achieve this objective is via development of speech recognition systems tuned for general deployment. For example, researchers at LIMSI (Paris) have found that a recognition system originally developed for a specific corpus is capable of doing reasonably well on other corpora as long as the basic range of speaking styles and dialects does not differ dramatically [30]. Such systems hold out the prospect of training acoustic models for phonetic classification in a language-independent manner. Although optimal performance still depends on manual tuning of the recognition system [30], portable recognition systems could provide the capability of performing a "first-pass" annotation automatically, with manual validation and correction performed afterwards to insure a high level of accuracy.

With respect to phonetic annotation this hybrid approach has been successfully deployed for a number of years. The ICSI transcription project [19] used the output of automatic alignments from the Johns Hopkins Switchboard automatic recognition system to generate phonetic-segment labels and boundaries, which were then (substantially) altered and adjusted by linguistically trained transcribers. The use of automatic labels saved a significant amount of time. A similar strategy has been used to annotate the phonetic component of the Spoken Dutch Corpus [6][7][12] as well as the Corpus of Spontaneous Japanese [26][35].

It is tempting to conclude from the experience of the ICSI, SDC and CSJ annotation projects that phonetic classification using automatic speech recognition alignment methods could be adapted to perform accurate phonetic labeling and segmentation without recourse to human intervention. However, a quantitative comparison of the labels and segmentation performed by humans and machines for a subset of the SWITCHBOARD corpus suggests that this objective is unrealistic in the absence of a computational interface between the word transcript and the phonetic labeler/segmenter (see the description of the MAUS system below). In terms of phonetic segmentation the machine-derived alignments deviated from manual segmentation by a mean of 32 ms across five separate recognition sites (the range was between 30 and 38 ms) [21]. The mean concordance between human transcribers for the same material was 8 ms. With respect to phonetic-segment labels, the aligners deviated from the output generated by human transcribers between 25 and 45% of the time (depending on the site and the criteria used for evaluating the accuracy of phonetic labeling). Under the most favorable evaluation conditions, at least one-quarter of the phonetic-segment labels generated by the automatic aligners differed from those annotated by trained human transcribers.

One method for dealing with such problems is to develop a "super aligner" customized to the pronunciation patterns encountered in a specific corpus (or language), as has been successfully applied to spontaneous German corpora by Schiel and colleagues in Munich. From an hour's worth of material that has been phonetically annotated in detail, it is possible to develop such an aligner incorporating knowledge of pronunciation variation both within and across words [4][5]. Such knowledge can be derived either from an elaborate set of rules or from statistical data [38]. The resulting system, MAUS (Munich Automatic Segmentation System), is capable of providing accurate phonetic labeling and segmentation for material comparable to what it was trained on, as long as a word transcript accompanies the acoustic signal. The concordance between human- and machine-generated labels and segmentation is very high [38], particularly when care is taken to tune the system to the patterns of pronunciation variation observed in the corpus [4][5].

However, a word transcript may not always be feasible to produce due to cost and labor constraints, nor does every site have the resources to develop a MAUS-like system to meet its specific corpus requirements. Under such circumstances, it would be useful to possess a tool capable of performing phonetic classification and segmentation automatically without recourse to a word transcript. This sort of "bottom-up" phonetic classifier would rely exclusively on acoustic properties of the signal (potentially supplemented by video recordings of the visible articulators) to ascertain the identity of phonetic segments and their associated boundaries. Such a system has been developed for spontaneous American English material using sophisticated neural networks trained on a phonetically labeled subset of the corpus (OGI Numbers) [10]. Its performance is as accurate as that of human transcriber. However, the classifier is confined to the phonetic segments contained in the corpus and therefore is not readily extensible to material from other languages (or even to different corpora from the same language).

4. Annotation of Phonetic Primitives

The acoustic-phonetic properties of all languages are grounded in the vocal tract. Elementary articulatory features (AFs), such as voicing, place and manner of articulation, apply to each and every language. How such features are phonetically realized and combine with other AFs vary from language to language, but the basic building blocks appear to be universal [24]. Rather than annotate a corpus in terms of phonetic segments, it is possible (in principal) to label the material at the AF level, in the hope that this form of annotation can be used more flexibly than one based exclusively on phones.

One advantage of AFs relative to phones is the smaller number of features per dimension. Although there are generally between 40 and 60 phones in a language, most AF dimensions contain only 2 to 6 features. Thus, AFs may provide a more tractable representation for machine classification, particularly under conditions of background noise (e.g., [28]).

Another potential advantage of AFs is their cross-linguistic potential. Wester and colleagues trained an AF classifier on an American English corpus (TIMIT) and used the system to classify AFs in Dutch [42]. Voicing and manner of articulation were both recognized nearly as well for Dutch as for English, but place of articulation did not transfer nearly as well [42]. This latter result raises the possibility that certain AF dimensions are realized similarly across languages, while others may be language specific and therefore require fine-tuning of the machine classifiers.

AFs may also be of utility for automatic annotation by virtue of their relation to such prosodic properties as syllable prominence and phrasal juncture. For example, voicing is controlled largely at the syllabic level and interacts with prosodic prominence with respect to its temporal dispersion within and across syllables and their phonetic constituents [20]. Place- and manner-of-articulation features, as well as certain vocalic parameters, also appear to be sensitive to prominence [23] and are likely to provide a more parsimonious framework for phonetic description than conventional phone-based systems.

The articulatory feature approach has been applied to both ASR and phonetic classification by a number of different groups – in English [11][13][14][27][34], German [28] and Dutch [42], and is likely to be used prominently in the future.

5. Automatic Segmentation of the Speech Signal

A principal issue that all acoustic classifiers must contend with is segmentation. Is a particular frame of speech the beginning of Word Y or the end of Word X (or somewhere in between)? Segmentation is sufficiently challenging from the technical perspective that current-generation ASR systems largely dispense with the problem via the use of hidden Markov models – a recognition system does not really "know" where it is in the speech stream, only that certain words are likely to proceed or follow the current point in time. This lexico-centric approach to recognition presents certain hazards when applied to phonetic characterization of speech because the articulatory realization of an utterance may transcend the concept of the "word" (see [32] for an example in Chinese). There is so much more to speaking than a mere sequence of words – the tone of voice, its emotional content, subtle nuances contained in the prosody, along with the visual information "invisible" to the audio recording.

Humans are far superior to machines with respect to segmentation of speech – at least at the syllabic level. Listeners quickly and accurately enumerate the number of syllables in a word or phrase, and can usually tell whether a syllable is "weak" or "strong" relative to its neighbors. Trained listeners can also tell how many phonetic constituents are contained in a word or syllable, though not always as accurately as they can specify syllables. Such knowledge, if provided to an automatic system, has the potential for significantly improving classification performance (e.g., [9][29]).

A number of different approaches have been taken to perform syllable segmentation automatically. It is possible to train neural networks to classify each frame of speech with respect to syllable onset, nucleus and coda components [39]. It is also possible to adopt a more principled signal-processing approach, as has been taken by Shire [44] or by Murthy and colleagues [36]. In each instance, approximately 90% of the syllables are currently segmented within a 40-ms tolerance window. A different strategy is to exploit knowledge of syllable structure, in concert with manner-of-articulation classifiers. Virtually all syllables possess a vocalic nucleus, and therefore it is possible to delineate syllables by virtue of their nuclei. This "vowel-spotting" technique has been applied to TIMIT sentences [40] as well as to SWITCHBOARD [9]. Segmentation at the syllabic level can be used, in concert with other AF classifiers (particularly manner of articulation), to delineate the phonetic constituents within a syllable with reasonable accuracy (as has been done for a portion of the SWITCHBOARD corpus [9]).

6. Prosodic Annotation of Spoken Corpora

Much of the phonetic, lexical, pragmatic and emotional content of speech is heavily influenced by its prosodic properties. In a recent study it has been shown that the identity and duration of phonetic segments in spontaneous English is highly correlated with syllable prominence [22][23], and it is likely that prosody impacts the phonetic characteristics of other languages as well. For such reasons it is important to annotate spoken corpora in terms of prosody, not just phonetics. A signal-driven syllable-prominence labeler (AutoSAL) has been developed to label spontaneous American English material. AutoSAL uses purely acoustic properties, such as vocalic duration, normalized energy and spectrum, to label syllable prominence in a manner comparable to that of a trained human transcriber [22].

Annotation at the supra-syllabic level (e.g., the phrase) is also important for capturing important information in the speech signal. ToBI is the most popular annotation system for capturing such detail, though it is unclear whether this classification system, which emphasizes pitch contours across and within syllables, is the optimum means of characterizing a language's prosodic patterns [43] except in instances where the language is explicitly tonal, such as Mandarin Chinese [32]. Its primary utility in non-tonal languages appears to be in demarcating phrasal boundaries [43]. A potentially more promising approach is to infer those portions of the acoustic signal consistently associated with specific prosodic events [16].

7. Automatic Annotation – Future Prospects

Spoken corpora of the future will be largely annotated through automatic methods, using a combination of sophisticated signal processing, statistical classifiers, and expert knowledge. It is important that such annotation tools be adaptable to a wide range of languages and speaking styles, as well as readily accessible to the speech research community around the world. Annotation at a fine-grained acoustic-phonetic level within a relatively theory-neutral framework (based on articulatory-acoustic features [8][20] or formant tracks [3]), that is melded to higher levels of linguistic annotation pertaining to prosody, emotion, and visual components of the speech signal, is most likely to provide the variety of empirical resources required to advance the state of speech technology and science.

Acknowledgements

I thank Francesco Cutugno, Li Deng, Sadaoki Furui, Lori Lamel, Kikuo Maekawa, Hema Murthy, T. Nagarajan, Florian Schiel, Klara Vicsi, Mirjam Wester, as well as the SALTMIL Executive Council, for their assistance in researching the background of this paper. I also thank Valerie Hazan for helpful comments on a preliminary draft. All opinions expressed are entirely those of the author and should not be construed as reflecting those of SALTMIL (http://193.2.100.60/SALTMIL) or any other organization.

References

[1] Allwood, J., Björnberg, M., Grönqvist, L, Ahls, E. and Ottesjö, C. "The Spoken Language Corpus at the Department of Linguistics, Göteborg University," Forum Qualitative Sozialforschung / Forum: Qualitative Social Research: http://www.qualitative-research.net/fqs-texte/3-00/3-00allwoodetal-e.htm.

[2] Aston, G. and Burnard, L. *The BNC Handbook.* Edinburgh: Edinburgh University Press, 1998. Also see the URL: http://www.hcu.ox.ac.uk/BNC/what/index.html.

[3] Bazzi, I., Acero, A. Deng, L. "An expectation-maximization approach for formant tracking using a parameter-free non-linear predictor," Proc. IEEE Int. Conf. Acoust. Speech Sig. Proc. (ICASSP), 2003.

[4] Beringer, N. and Neff, M. "Regional pronunciation variants for automatic segmentation," Proc. Int. Conf. Spoken Lang. Proc., Beijing, 2000.

[5] Beringer, N. and Schiel, F. "The quality of multilingual automatic segmentation using German MAUS," Proc. Int. Conf. Spoken Lang. Proc., pp. 728-731, 2000.

[6] Binnenpoorte, D., Goddijn, S. and Cucchiarini, C. "How to improve human and machine transcriptions of spontaneous speech," ISCA/IEEE Workshop on Spont.

Speech Proc. Recog., pp. 147-150, 2003.

[7] Boves, L. and Oostdijk, N. “Spontaneous speech in the spoken Dutch corpus,” ISCA/IEEE Workshop on Spont. Speech Proc. Recog., pp. 171-174, 2003. See the URL: http://lands.let.kun.nl/cgn/ehome.htm

[8] Carson-Berndsen, J. “Multilingual time maps; Portable phonotactic models for speech technology applications,” SALTMIL Workshop, LREC Conference, Las Palmas, 2002.

[9] Chang, S. *A Syllable, Articulatory-Feature, and Stress-Accent Model of Speech Recognition.* Ph.D. Thesis, University of California, Berkeley. ICSI Technical Report TR-02-007.

[10] Chang, S., Shastri, L. and Greenberg, S. “Automatic phonetic transcription of spontaneous speech (American English),” Proc. 6th Int. Conf. Spoken Lang. Proc., Beijing, 2000.

[11] Chang, S., Greenberg, S. and Wester, M. “An elitist approach to articulatory-acoustic feature classification,” pp. 1725-1228, 2001.

[12] Demuynck, K., Laureys, T. and Gillis, S. “Automatic generation of phonetic transcriptions for large speech corpora,” Proc. Int. Conf. Spoken Lang. Proc., pp. 333-336, 2002.

[13] Deng, L. and Sun, D. “A statistical approach to automatic speech recognition using the atomic speech units constructed from overlapping articulatory features,” J. Acoust. Soc. Am. 95: 2702-2719, 1994.

[14] Espy-Wilson, C. “A Feature-Based Approach to Speech Recognition,” J. Acoust. Soc. Am. 96: 65-72, 1994.

[15] European Language Resources Association (ELRA) – www.icp.inpg.fr/ELRA

[16] Ferrer, L., Shriberg, E. and Stolcke, A. “A prosody-based approach to end-of-utterance detection that does not require speech recognition,” Proc. IEEE Int. Conf. Acoustics, Speech Sig. Proc, 2003.

[17] French, N. R., Carter, C. W. and Koenig, W., “The words and sounds of telephone conversations,” Bell System Tech. J. 9: 290-324, 1930.

[18] Godfrey, J.J., Holliman, E.C. and McDaniel, J. “SWITCHBOARD: Telephone speech corpus for research and development.” Proc. IEEE Int. Conf. Acoust. Speech Sig. Proc. (ICASSP-92), pp. 517-520, 1992.

[19] Greenberg, S. “Speaking in shorthand - A syllable-centric perspective for understanding pronunciation variation,” Speech Comm. 29: 159-176, 1999. See the URL: http://www.icsi.berkeley.edu/real/stp/index.html

[20] Greenberg, S. “Pronunciation is key to understanding spoken language,” Proc. 15th Int. Cong. Phon. Sci., 2003.

[21] Greenberg, S. and Chang, S. “Linguistic dissection of switchboard-corpus automatic speech recognition systems,” Proc. ISCA Workshop on Automatic Speech Recognition: Challenges for the New Millennium, pp. 195-202, 2000.

[22] Greenberg, S., Chang, S. and Hitchcock, L. “The relation between stress accent and vocalic identity in spontaneous American English discourse,” Proc. ISCA Workshop Prosody Speech Recog. Understanding, pp. 51-56, 2001. See the URL: www.ICSI.Berkeley.EDU/~steveng/prosody/index.html

[23] Greenberg, S., Carvey, H., Hitchcock, L. and Chang, S. “Beyond the phoneme – A juncture-accent model for spoken language,” Proc. 2nd Int. Hum. Lang. Tech. Conf., pp. 36-43, 2002.

[24] Jakobson, R., Fant, G. and Halle, M. *Preliminaries to Speech Analysis: The Distinctive Features and their Correlates.* Cambridge: MIT Press, 1963 [originally issued as an MIT technical report in 1952]

[25] The Kiel Corpus. Kiel University, Germany. http://www.ipds.uni-kiel.de/forschung/kielcorpus.en.html.

[26] Kikuchi, H and Maekawa, K. “Performance of segmental and prosodic labeling of spontaneous speech,” ISCA/IEEE Workshop on Spont. Speech Proc. Recog., pp. 191-194, 2003.

[27] King, S. and Taylor, P. “Detection of phonological features in continuous speech using neural networks,” Comp. Speech Lang. 4: 333-345, 2000.

[28] Kirchhoff, K., Fink, G.A. and Sagerer, G. Combining acoustic and articulatory feature information for robust speech recognition. Speech Comm 37: 303-319, 2002.

[29] Konig, Y. and Morgan, N. “Modeling the dynamics in connectionist speech recogition – the time index model,” Proc. Int. Conf. Sp. Lang. Proc., 1994.

[30] Lamel, L. “Some Issues in Speech Recognizer Portability,” SALTMIL Workshop, LREC Conference, Las Palmas, 2002.

[31] Lander, T., Cole, R.A., Oshika, B.T. and Noel, M. “The OGI 22 language telephone speech corpus,” Proc. 4th European Conf. Speech Comm. Tech. (Eurospeech-95), 1995. http://cslu.cse.ogi.edu/corpora/corpCurrent.html.

[32] Lee, L.-S., Ho, Y., Chen, J.-F., Chen, S.-C. “Why is the special structure of the language important for Chinese spoken language processing?” Proc. 8th European Conf. Speech Comm. Tech. (Eurospeech-2003), 2003.

[33] Linguistic Data Consortium – www.ldc.upenn.edu

[34] Liu, S.A. “Landmark detection for distinctive feature-based speech recognition,” J. Acoust. Soc. Am. 100: 3417-3430, 1996.

[35] Maekawa, K. “Corpus of spontaneous Japanese: Its design and evaluation,” ISCA/IEEE Workshop on Spont. Speech Proc. Recog., pp. 7-12, 2003. See the URL: http://www2.kokken.go.jp/~csj/public/

[36] Nagarajan, T., Murthy, H.A. and Hegde, R.M. “Group delay based segmentation of spontaneous speech into syllable-like units. ISCA/IEEE Workshop on Spont. Speech Proc. Recog., pp. 115-118, 2003.

[37] Schiel, F. “Speech and speech-related resources at BAS,” Proc. 1st Int. Conf. Lang. Res. Eval. (LREC), pp. 343-349, 1998. See the URL: http://www.phonetik.uni-muenchen.de/Bas/BasHomeeng.html

[38] Schiel, F. “Automatic phonetic transcription of non-prompted speech,” Proc. 14th Int. Cong. Phon. Sci., pp. 607-610, 1999. See the URL: http://www.phonetik.uni-muenchen.de/Forschung/Verbmobil/VM14.7eng.html.

[39] Shastri, L. Chang, S. and Greenberg, S. “Syllable detection and segmentation using temporal flow neural networks,” Proc. 14th Int. Cong. Phon. Sci., pp. 1721-1724, 1999.

[40] Sirigos, J., Fakotaki, N., Kokkinakis, G. “A hybrid syllable recognition system based on vowel spotting,” Speech Comm 38: 427-440, 2002.

[41] Svartvik, J. (ed.) *The London Corpus of Spoken English: Description and Research.* Lund Studies in English 82, Lund University Press, 1990. See the URL: http://khnt.hit.uib.no/icame/manuals/LONDLUND/INDEX.HTM for additional information.

[42] Wester, M., Greenberg, S. and Chang, S. “A Dutch treatment of an elitist approach to articulatory-acoustic feature extraction,” Proc. 7th European Conf. Speech Comm. Tech. (Eurospeech-2001), pp. 1729-1732, 2001.

[43] Wightman, C. “ToBI or not ToBI,” Proc. Speech Prosody-2002, 2002.

[44] Wu, S-L., Shire, M., Greenberg, S. and Morgan, N. “Integrating syllable boundary information into speech recognition,” Proc. IEEE Int. Conf. Acoust. Speech Sig. Proc. (ICASSP), pp. 987-990, 1997.