NEURAL NETWORK MOTIVATION FOR SEGMENTAL DISTRIBUTION

Eric Keller

LAIP - Lettres, University of Lausanne, 1015 Lausanne, Switzerland, Eric.Keller@imm.unil.ch

ABSTRACT

Feature representations mediating between acoustic input and symbolic representation promise to reduce learning time needed for automatic speech signal segmentation. Experiments are reported that circumscribe acoustic inputs and appropriate feature sets for neural network (NN) training.

1. INTRODUCTION

The phonetic sciences are centrally concerned with the relationship between linguistic and acoustic representations of speech. This generally means establishing systematic and explicit relationships between the elements of a phonetic chain and corresponding acoustic segments in a speech signal.

Despite considerable advances in understanding articulatory and acoustic relationships mediating between these levels of representation, the structure required for inducing the phonetic representation from a given speech signal remains in part indeterminate (a given symbol can be represented by different acoustic states, and vice versa), and in other respects obscure (a symbol can not always be related to identifiable acoustic information, or vice versa). This has proven a handicap in speech recognition (the "phoneme recognition problem") as well as in segmenting and labelling the speech signal automatically (the "phoneme matching problem"), a handicap that has traditionally been overcome by Markovian statistical modelling of LPC or cepstral coefficients derived from the signal. On the basis of substantial learning and by taking into account preceding and current acoustic information, fairly reliable predictions can be made concerning the likely phonetic representation of the acoustical segment under consideration.

The information required for statistical identification can be considerable, since it is proportional to the number of states to be distinguished. Other parameters being held constant, if a (mythical) language distinguished only two phonemes, relevant acoustic information could easily be related to phonetic symbols. But solving the identification or matching problem in typical human languages implies learning many more states and constraints between states, which in turn leads to considerable delays between the onset of learning and the point where the recognition or matching problems can begin to be solved with any degree of success.

Improvements can be expected if an intermediate representation between the acoustical signal and the phonetic symbol can be assumed, i.e., if feature representations can be found to reliably mediate between symbolic and acoustic states. The learning problem can then be simplified, and perhaps more importantly, be explicated in terms of more direct — and hopefully more intuitively comprehensible — relationships between signal and features. This is the approach we pursued in a NN-driven speech segmentation technique. First results are presented here of experiments performed to define an NN design and a phonetically explicit feature matrix for English and French.

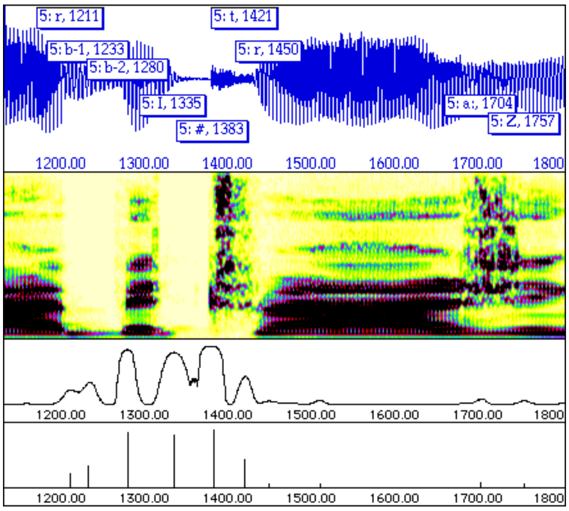

Figure 1. Segmentation by identification of points of maximal spectral change. Top: segment of signal taken from the word "arbitrage" /-rbItraZ-/. 2nd from top: wide-band spectrogram. 3rd from top: spectral changes. Bottom: Points of maximal spectral change.

2. METHOD

2.1. The Segmentation Procedure

The segmentation procedure proceeds in five steps:

(1) *Segmentation.* 512-point FFT spectra obtained from pre-emphasis and hamming window-processed signals are calculated at 1-ms intervals for the entire signal, to produce the numeric equivalent of a narrow-band spectrogram. A 30-40 ms window (depending on speech rate) is passed over the spectrogram, and auditorily weighted average vectors are calculated for the left and right halves of the window. At the window's midpoint, cumulative squared and amplitude-weighted differences between the left and the right halves of the window are calculated over all spectral rays. These differences identify points of maximal change in the spectrogram and serve to establish potential points of segmentation (Figure 1).

(2) *Acoustic parameters.* Input parameters were derived directly from the signal as well as from a non-uniform filter bank based on the spectrogram and defined with eleven non-overlapping spectral bands (cf. [1] p. 91). Bands are delimited by the first 22 bands of the Bark scale, where each spectral band corresponds to two Bark scale bands. (Originally, a

more traditional 22-band filter based on the Bark scale was used, but was found to exhibit insufficient generalisability in steps 3 and 4.) On this basis, the following parameters were identified for each segment: (a) duration, (b) global amplitude, (c) global amplitude slope, (d) onset amplitude slope, (e) average f0, (f) f0 slope; plus for each spectral band: (g) average amplitude, (h) overall amplitude slope and (i) onset amplitude slope. All acoustic parameters are standardised to the ±1.0 range, with reference to the parameter's values over the entire spectrogram.

(3) *Training.* On the basis of a minimum of 500 hand-verified and phonetically labelled segments, relationships between the acoustic parameters and 23 binary phonetic features (see below) are learned by a Quickprop algorithm (essentially, the BP program by Don Tweter, available from http://www.unidial.com/~drt). The current base configuration defines 7 hidden nodes, a number arrived at by observing the network's learning behaviour (see below) as well as by a principal components analysis. Learning ends when the 0.999 criterion is reached. for both data and bootstrap sets.

(4) *Testing.* For a new signal, speech segments are identified by step 1, and hypothetical binary features are calculated from the acoustic parameters derived in step 2 and the weights calculated in step 3. From the hypothetical features, hypothetical segment labels are constituted.

(5) *Matching.* Time-warping relates expected feature bundles derived from the phonetic symbols to the hypothetical feature bundles obtained in steps 1-4. Expected durations are calculated for phonetic segments, and a matching process attempts to infer the phonetic symbol associated with each empirically determined segment.

2.2. Experiments in NN Design

Experiments were performed to arrive at a first stable NN configuration. "Learnability" was defined as iterations to criterion averaged over the set of 23 features and in terms of the maximum number of iterations. The network halts learning after 4000 iterations, at which point learning was considered impossible under the given configuration.

The base configuration as described above and a feature matrix as defined below were assumed. Only the parameter(s) under consideration was/were manipulated. The base NN was a fully-connected quickprop network with 39 acoustic input nodes, seven hidden nodes, and binary output nodes. Each of the 23 features was trained separately.

Training data consisted of 1000 English and 1000 French segments in distinctly articulated individual words and short phrases, recorded from a talented multilingual female speaker. The corpus was segmented using the procedure described above and segmentations were manually verified. Experiments were run primarily on the English data, except for the language manipulation noted below (Table I).

The number of hidden nodes was varied from 10 to 5, in order to determine the lowest number of nodes capable of reliable learning. Lower numbers of nodes are favoured in order to further the network's generalisation capacity. A configuration with seven hidden nodes showed particularly good performance in comparison with configurations with either more or fewer hidden nodes. Learning proved impossible at five or fewer nodes.

Acoustic input parameters. Preliminary analyses had suggested that each acoustic parameter of the base configuration had independent contributing value. By removing each parameter individually, this was confirmed. In particular, all three types of spectral band information contributed crucially to learnability: relative amplitude, overall amplitude slope and onset amplitude slope. The general slope designates the ratio between averaged relative amplitudes in the initial third and the last third of the segment's measures. Onset slope is the ratio between of the first and the second sixths of such measures. Learning proved impossible when each of these parameters served as the sole input to the NN, or even when only two parameters were combined. All three parameters were required to permit learning in the base configuration.

Language Generalisation. The base configuration for English was also tested on the French data set (Table I). The NN learned data from French at least as easily as from English.

TABLE I: NN MANIPULATIONS

	Average number of iterations	Ratio to base configu-ration	Maximum number of iterations
BASE CONFIGURATION	289	1.0	803
HIDDEN NODES			
10	213	0.74	711
9	236	0.82	563
8	306	1.06	1997
7	289	1.00	803
6	414	1.43	1267
5	825	2.85	>4000
ACOUSTIC INPUT			
Amplitude only	2847	9.85	>4000
Overall slope only	2624	9.08	>4000
Onset slope only	2625	9.08	>4000
Amplitude + overall slope	1249	4.32	>4000
Amplitude + onset slope	1208	4.18	>4000
Overall slope + onset slope	886	3.07	>4000
Amplitude + overall slope + onset slope	289	1.00	803
All except duration	520	1.80	>4000
All except avg. f0	349	1.21	1048
All except f0 slope	394	1.36	1288
All except global amplitude	333	1.15	1031
All except onset amp. slope	434	1.50	1787
All except global amp. slope	359	1.24	1038
LANGUAGE			
English	289	1.00	803
French	206	0.71	413

2.3. Current Deficiencies

At each step, the procedure is exposed to errors. Current work is directed at identifying and suppressing such errors through further development, iterative experiments and adjustments. The main areas of investigation are the following:

(1) *Segmentation errors.* In addition to identifying "real" segment boundaries, the procedure also identifies phonetic changes that are not related to a symbolic segment boundary. This tends to occur at periods of rapid formant change, in the middle of fricatives, or at slow transitions between sounds, where the spectrogram exhibits periods of little acoustic energy. Also, expected segment boundaries are sometimes not identified in the acoustic material, because the two adjoining sounds are acoustically similar (e.g., "re-invent") or because articulation is indistinct. Currently, such errors are corrected manually. An automatised correction procedure is envisageable.

(2) *Feature identification errors.* Some (e.g., vowel) features depend on acoustic information that is contextually dependent and is thus not structured in the same manner as acoustic information for the same feature in another segment. In such cases, the NN often identifies an erroneous segment. Currently, many such errors are corrected in the matching step (5). Further improvements are expected through the use of contextual information.

(3) *Matching errors.* The results of the time warp are affected by the precision of durations predicted for the transcribed phonetic string. If segments are predicted much before or after the time that they are actually found in the signal, correct matching is nearly impossible. The current model predicts durations on the basis of average verified segment durations which limits matching performance to single words or word groups. Future work will be directed at improving the durational prediction.

3. THE FEATURE MATRIX

3.1. Defining the matrix

Performance in the current system depends largely on the phonetic feature matrix employed. Features that are difficult to learn show low generalisability, while those that are easily learned tend to show greater reliability in generalisation. Also, some features are easier to link to acoustic input than others. Finally and in contrast to traditional phonetic teaching, a matrix consisting of fewer features is not necessarily better for segmentation than a matrix that employs many features, since a large, non-redundant feature matrix offers more distinctiveness than a smaller, equally non-redundant matrix.

Over a lengthy (but non-exhaustive) set of experiments, the link between features and spectral information was optimised. Base NN configuration and training procedures as described above were used. Feature definitions for input segments were varied, and a redefinition of features was attempted if the association between acoustic input data and binary feature could not be learned in 1000 iterations or less. Subsequent to each redefinition, the network was retrained. Redefinition and retraining continued until three conditions were met: (1) each phonetic symbol in the English and French symbol set was uniquely defined, (2) English and French feature definitions for comparable sounds were identical, (3) each feature could be learned in less than 1000 iterations.

3.2. Results of matrix manipulations

The following principles emerged (Figure 2):

1. *Scope.* A few features are relevant to the entire segment inventory, while most are relevant only to certain groups of sounds. Voicing and nasality can be successfully distinguished in both consonants and vowels (nasality for vowels was only tested in French). These features depend on similar acoustic features in all segments, and their articulatory source is the rear of the vocal tract. By contrast, most other features are specific to either the vocalic or the consonantal domain. For example, the feature FRONT proved well-nigh impossible to learn (3200 iterations) when defined with distinctive values for both vowels and consonants. When "frontness" was redefined as the feature FRONT for vowels and as ALVEOLAR for consonants, the association could be learned easily (less than 500 iterations each).

As a consequence of this experimentation, three groups of sounds were distinguished: vowels + semivowels, non-fricative consonants, and fricative consonants. The status of certain sounds is language-specific. French fricative /r/ is best grouped with the fricative consonants, while the English rolled /r/ is best associated with the semi-vowels.

2. *Markedness.* Most group-specific features behave as "identifiers" or "markers". For example, the feature +HIGH identifies the vowels /i:/ and /u:/. All other sounds (vowels as well as consonants) are marked -HIGH. This type of specification is learned easily, and it suggests that acoustic information in specific bands (e.g., for HIGH, relatively high energy in the below-1 kHz bands) are the relevant predictors of the marked features. Absence of this type of acoustic information is interpreted by the network as an absence of marking, and thus permits correct switching of the binary feature.

3. *Groups of feature states.* Many binary features are best defined as mutually exclusive states in *groups* of states. For example, high-low or front-back distinctions for vowels are best defined in terms of mutually exclusive feature "activations" (e.g, FRONT - CENTRE - BACK, or HIGH - OPEN - MID - LOW). In other systems, such features would be considered to be multivalued (e.g., HEIGHT 1-4). Some other features, on the other hand, are clearly binary (e.g., VOICED, NASAL). A special case is presented by *additive* features relating to fricative sounds. For sounds such as /S/ or /Z/, the frication band extends upwards from lower points in the spectrum than that for sounds such as /s/ and /z/. Yet in the higher parts of the spectrum (e.g., above about 4 kHz), *both* groups show strong fricative noise. Consequently, non-exclusive +SHARP features states are specified for all members of the fricative set, while only /S/ and /Z/ are additionally marked by a +STRIDENT feature. The fricative feature +GRAVE is reserved for the English sounds /T/ and /D/ which exhibit quite a different fricative behaviour. The sound /h/, finally, is +FRICATIVE but receives no further marking, in view of its generally weak acoustic frication.

4. CONCLUSION

In experiments involving NN learning of relations between acoustic input and phonetic features, a stable NN configuration was identified. It consists of a fully connected NN with inputs from 11 spectral bands, f0, amplitude and duration, 7 hidden nodes and output nodes for 23 binary features.

Design experiments showed that relations between simple spectral parameters and features could be learned in few iterations (accuracy has not been assessed systematically yet-). Of particular surprise was the crucial contribution of *slope* information (overall slopes as well as onset slopes in spectral band output, f0 and global amplitude). This information presumably codes a segment's dynamic features [2]. Further experiments must establish if this design is superior or inferior to the more common LPC- or cepstrum-based preprocessing designs [1, for recent work, see e.g. 3].

The matrix manipulation experiments led to stable and compatible solutions for English and French, suggesting that a segmentation system of this type can easily be expanded to handle multiple languages. Additional performance gains are expected from taking into account preceding and succeeding acoustic information. Further testing will be directed at a greater number of subjects and speech styles, and at fully permuting the feature matrix in a stepwise procedure.

5. ACKNOWLEDGEMENTS

Supported by grants from the Swiss Confederation under the COST 258 and the KTI/CDI programmes. Grateful acknowledgement to B. Pfister, C. Traber (TIK-ETH, Zurich), B. Zellner (LAIP) and to K. Nadir (Swisscom, Berne) for collaborative work in constituting the multilingual database under a contract from Swisscom.

6. REFERENCES

1. Rabiner, L., & Juang, B.-H. (1993). *Fundamentals of Speech Recognition.* Prentice-Hall.
2. Olive, J.P., Greenwood, A., & Coleman, J. (1993). *Acoustics of American English Speech: A Dynamic Approach.* Springer-Verlag.
3. Mazin Rahim, M., Yoshua Bengio, Y., & and Yann LeCun, Y. Discriminative feature and model design for automatic speech recognition. *Proceedings of EUROSPEECH '97, 1,* 75 - 78.

IPA	ASCII	segmt	vocal	voicd	long	front	centr	back	round	high	open	mid	low	strss	stop	labl	alvlr	paltl	velar	nasal	fric	strid	sharp	grave
¨i	'i	+	+	+	+	+	—	—	—	+	—	—	—	+	—	—	—	—	—	—	—	—	—	—
¨I	'I	+	+	+	—	+	—	—	—	—	+	—	—	+	—	—	—	—	—	—	—	—	—	—
¨e	'e	+	+	+	+	+	—	—	—	—	+	—	—	+	—	—	—	—	—	—	—	—	—	—
¨E	'E	+	+	+	—	+	—	—	—	—	—	+	—	+	—	—	—	—	—	—	—	—	—	—
¨Q	'&	+	+	+	—	—	+	—	—	—	—	+	—	+	—	—	—	—	—	—	—	—	—	—
¨«	'*	+	+	+	—	—	+	—	—	—	+	—	—	+	—	—	—	—	—	—	—	—	—	—
¨a	'a	+	+	+	—	—	+	—	—	—	—	—	+	+	—	—	—	—	—	—	—	—	—	—
¨A	'A	+	+	+	+	—	—	+	—	—	—	—	+	+	—	—	—	—	—	—	—	—	—	—
¨	'O	+	+	+	+	—	—	+	+	—	—	+	—	+	—	—	—	—	—	—	—	—	—	—
¨o	'o	+	+	+	—	—	—	+	+	—	+	—	—	+	—	—	—	—	—	—	—	—	—	—
¨U	'U	+	+	+	—	—	+	—	+	—	+	—	—	+	—	—	—	—	—	—	—	—	—	—
¨u	'u	+	+	+	+	—	—	+	+	+	—	—	—	+	—	—	—	—	—	—	—	—	—	—
w	w	+	+	+	—	—	—	—	+	+	—	—	—	—	—	—	—	—	—	—	—	—	—	—
j	j	+	+	+	—	—	—	—	—	+	—	—	—	—	—	—	—	—	—	—	—	—	—	—
l	L	+	+	+	—	—	—	—	—	+	—	—	—	—	—	—	+	—	—	—	—	—	—	—
r	r	+	+	+	—	—	—	+	—	+	—	—	—	—	—	—	—	+	—	—	—	—	—	—
p	p	+	—	—	—	—	—	—	—	—	—	—	—	—	+	+	—	—	—	—	—	—	—	—
b	b	+	—	+	—	—	—	—	—	—	—	—	—	—	+	+	—	—	—	—	—	—	—	—
f	f	+	—	—	—	—	—	—	—	—	—	—	—	—	—	+	—	—	—	—	+	—	—	—
v	v	+	—	+	—	—	—	—	—	—	—	—	—	—	—	+	—	—	—	—	+	—	—	—
m	m	+	—	+	—	—	—	—	—	—	—	—	—	—	—	+	—	—	—	+	—	—	—	—
t	t	+	—	—	—	—	—	—	—	—	—	—	—	—	+	—	+	—	—	—	—	—	—	—
d	d	+	—	+	—	—	—	—	—	—	—	—	—	—	+	—	+	—	—	—	—	—	—	—
T	T	+	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	+	—	—	+
D	D	+	—	+	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	+	—	—	+
s	s	+	—	—	—	—	—	—	—	—	—	—	—	—	—	—	+	—	—	—	+	—	+	—
z	z	+	—	+	—	—	—	—	—	—	—	—	—	—	—	—	+	—	—	—	+	—	+	—
S	S	+	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	+	—	—	+	+	+	—
Z	Z	+	—	+	—	—	—	—	—	—	—	—	—	—	—	—	—	+	—	—	+	+	+	—
n	n	+	—	+	—	—	—	—	—	—	—	—	—	—	—	—	—	+	—	+	—	—	—	—
k	k	+	—	—	—	—	—	—	—	—	—	—	—	—	+	—	—	—	+	—	—	—	—	—
g	g	+	—	+	—	—	—	—	—	—	—	—	—	—	+	—	—	—	—	—	—	—	—	—
h	h	+	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	+	—	—	—
N	G	+	—	+	—	—	—	—	—	—	—	—	—	—	—	—	—	—	+	+	—	—	—	—
x	x	+	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	+	—	+	+	+	+
?	q	+	—	—	—	—	—	—	—	—	—	—	—	—	+	—	—	—	—	—	—	—	—	—
V	V	+	—	+	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
#	#	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—

Figure 2. Current feature chart for English. Unstressed versions of vowels are not shown. The symbols [V] and [#] designate preburst periods of voiced and unvoiced stops respectively. Experimentation still in progress. [Note: because of a limitation in Microsoft Word, the postscript version of this article does not show the phonetic symbols correctly. Please refer to the ASCII equivalents.]