AUTOMATIC EXTRACTION OF PHONETIC FEATURES IN SPEECH, USING NEURAL NETWORKS.

F. BIMBOT, G. CHOLLET, J.P. TUBACH.

Télécom Paris - Département SIGNAL, C.N.R.S. - URA 820.

ABSTRACT

We report in this paper a series of experiments aiming at automatically extracting phonetic features in speech, using a specific family of neural networks, namely TDNNs. The results show the interest that exists in using phonetic knowledge to guide speech recognition. The visualisation of connection weights inside the optimised networks illustrates the various strategies used by TDNNs for classifying sounds into features, after their acoustic content

RESUME

Cet article présente un ensemble d'expériences visant à l'extraction de traits phonétiques à partir de signal de parole, à l'aide d'une famille particulière de réseaux neuro-mimétiques : les TDNNs. Les résultats montrent l'intérêt d'utiliser des connaissances phonétiques pour orienter la tâche des réseaux, en reconnaissance de parole. La visualisation des poids des connexions dans les TDNNs après optimisation illustre la façon dont ceux-ci utilisent les caractéristiques acoustiques des sons pour déterminer les différents traits.

INTRODUCTION

Connectionist networks are one of the possible tools for performing classification tasks, such as those required in particular for speech recognition. Multi-layer neural networks are made of several layers of cells (or nodes), each of them delivering, as an output, a (usually) non-linear transform of the input; for instance a sigmoid. The input itself is obtained as the weighted sum of the activations of the nodes in the previous layer that are connected to the current node. A multi-layer neural network can be viewed as a black box made of a large number of elementary units with a rather simple individual function, the global behaviour of which makes it possible to model quite complicated non-linear transfer functions.

The number of nodes in each layer and the connection structure (i.e. what is called the architecture) is usually fixed a priori, whereas the values of each weight (what could be called the "furniture") are task specific and classically estimated by the back-propagation algorithm, given a set of training examples. In other words, it is possible, with neural networks, to automatically learn some non-linear discriminations between families of patterns, without any careful human time-consuming specification of classification rules. However, the computing time required is rather high.

THE TDNN

Waibel et al introduced TDNNs (Time-Delay Neural Networks) as a specific neural network architecture that can take into account the "dynamic nature of speech" [1]. Indeed, such a network is able to represent temporal relationships between successive acoustic time-slice (frames), which is a property of major importance, since static characteristics of the speech signal are certainly unsufficient for a proper identification of sounds or sound features. Moreover, TDNNs provide some invariance under time-translation, which lessens the sensitivity of the processing in front of unavoidable segmentation inaccuracies. Figure 1 illustrates the TDNN structure.

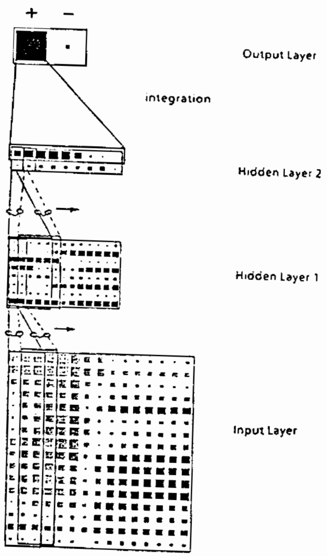

Figure 1 : A TDNN (adapted from Waibel).

TDNNs have 4 layers (an input layer, 2 hidden layers and an output layer). Successive layers are sparsely connected (in practice, some connections weights are set to zero). Moreover, there exist groups of weights that are constrained to be identical to one another, in order to warrant the time-shift invariance property. More detailed descriptions of TDNNs can be found in several papers [1] [2] [3].

In recent publications, TDNNs proved to be very efficient for tasks such as classifying [b] vs [d] vs [g] among a set of observation containing phonetic realisations of these 3 phonemes, in japanese [1], or in french [4]. Waibel et al. also showed that a modular all-phoneme classification network could be successfully designed on the basis of the collaboration of a collection of such elementary networks (one for [bdg], one for [ptk], one for [fsʃ], ...), provided an other network makes concurrently a rough classification of each sound, in order to decide which of these elementary networks is to be resorted to. In parallel to these developments, Haffner et al. proposed fast learning methods for TDNNs that bring down the training time to reasonable figures [2].

All these properties make of TDNNs appealing tools for classifying speech sounds, with the ultimate goal of speech recognition. Our approach of phonetic features extraction is slightly different from the modular all-phoneme recognition process proposed by Waibel et al. For the latter, a hierarchical decision is needed, whereas, for the former, several parallel classifications are made and then integrated in a final phoneme identification.

PHONETIC FEATURES

The phonetic description of speech events is classically based on distinctive features. Features are usually binary, since they indicate the presence or absence of a specific characteristic for a given sound (or family of sounds).

Features can be defined according to acoustical, articulatory, or even linguistic properties. For instance, the *grave / acute* opposition is based on acoustical considerations, while *front / back* is articulatory and *vowel / consonant* based on higher-level linguistic concepts.

Each phoneme (or, more precisely in our case, each phone) results from the simultaneous realisation of several elementary binary features, which Trubetzkoï describes as a "Korrelazions-bündel" (i.e. a bundel of correlations), and which Jakobson qualifies as a "bundle of concurrent binary distinctive features" [5].

Several systems of features have been proposed, among which must be mentioned those of Malmberg [6] and Rossi [7] for the french language, those of Jakobson [5] and Chomsky [8], in english. Jakobson's classification was the first one to rely on acoustical criteria only.

A few speech recognition systems have been calling for feature extraction (or macro-classes identification) strategies, most of them using expert systems that track cues on the acoustic signal, with the help of rules.

The approach using neural networks can be understood as a mean of expressing and modeling, with a mathematical non-linear tool, the (sometimes) complex relationship existing between the abstract notion of phonetic feature and its physical manifestation through acoustic cues.

PROTOCOL AND CORPUS

We have been evaluating the use of TDNNs for different feature extraction tasks, for the french language.

A set of 13 binary features was thus designed, relying on the most classical phonetic oppositions. This set allows a total description of the french phonetic system. It is however not minimal, but our goal was to investigate what type of oppositions TDNNs are best adapted to.

An other set of 6 discriminative and minimal "random" features was artificially built, so that it opposes 34 phonemes with one another. This set has, of course, no phonetic background, since phonemes that have "nothing to do" with each other are member of the same random class and opposed to an other (complementary) class of phonemes that have "nothing to do" with each other either ! These artificial features however serve as a point of comparison for judging the relevance and the usefulness of classical phonetic oppositions versus arbitrary ones.

Other sets or "ternary" features have also been used [3], but results are not reported in this article.

The corpus for the experiments contains 200 french phonetically balanced sentences [9], uttered by one male speaker in excellent recording conditions and sampled at 16 kHz. Each phoneme realisation was coarsely labeled at the center by hand, and represented by 16 time frames (8 on each side of the label) of 16 Mel-scaled filter-bank coefficients; in other words, a quite low resolution (256 pixels) spectrographic representation, with non-linear frequency scale. The corpus contains 5270 tokens (unbalancedly dispatched between the phonetic classes), which were halfed between a learning set and a test set.

The corpus was transcribed prior to its production and labeled according to this normative transcription, using 34 symbols :

i	y					u				#		
	e	ø			o							
		ɛ	œ	ɔ				ɛ̃				ɔ̃
			a							ã		
p	t	k		b	d	g	j	ɥ				w
f	s	ʃ		v	z	ʒ						
				m	n	ɲ			l		ʁ	

The phonetic system used in our experiments. [#] denotes "silence". [ə] = [œ].

Automatic phonological rules were applied to modify the features of some phonemes, to account for basic contamination effects.

Evaluations were done on the whole corpus, on the sub-corpus of vowels only (2952 items) and on the sub-corpus of consonants only (2318 items). Sub-corpuses were also halfed in training and test sets.

The cross-validation strategy was used to decide when to stop the learning phase [3].

RESULTS

Performances on the whole corpus, on the vowels-only corpus and on the consonants-only corpus, for each feature, are shown in Table 1 below. Two scores are given for each experiment : one (in bold) is the score of correct feature identification on the test set (i.e. for the the data that were not used to train the network), while the second score (in standard characters) corresponds to the score obtained on the training set itself (self-consistency). Naturally, the second score is usually higher than the first one, and the difference between the two gives a hint of the generalisation capability of the network; that is its ability to solve a similar problem to the one it was trained for, with data that it has not seen yet.

On the whole corpus, all scores (but one) range between 90 % and 99 %, with more than half of them over 95 %. Moreover, all scores for phonetic features (but one) are higher (by approximately 10 %, in the average) than the average score for arbitrary random features. This clearly evidences the contribution of phonetic knowledge for determining what kind of task the TDNN is most likely to work with. Beside this, while features related to manner of articulation are very efficiently detected (features III, V, XI, XII, XIII), those linked to the more abstract notion of place of articulation provide less satisfactory results.

Not all phonetic features that were tested on the whole corpus were also experimented for the vowel and the consonant sub-corpuses, but only some of them that allow a discriminative non-redundant system. The improvement of feature extraction when moving from the whole corpus to the consonants-only corpus is rather disappointing, but must be owed to a change of the repartitions. Conversely, scores for vowels-only improve significantly, in general. In both cases, the self-consistency tends to increase, since the number of learning examples is smaller and makes it possible for the TDNN to memorise the particularities of the training data (which is clearly undesirable).

VISUAL EXPLORATION

It can be shown that, under certain constraints, the matrices of weights within the first layer of TDNNs can be viewed as typical patterns that are searched for in the spectral picture of the input token to classify [10]. In other words, TDNNs develop in some ways their own expertise for classifying speech sounds, a little bit like a human expert would do, from experience.

In figure 2 (last page), we have visualised 3 sets of weight matrices, for 3 features : *voiced / unvoiced*, *nasal / non-nasal*, and *vowel / consonant*. A full comment is given with the figure : the cues used by the network are most of the time in accordance with the classical acoustic descriptions of phonetics, and thus directly interpretable, which was not at all a priori warranted.

CONCLUSION

Phonetic knowledge can thus be used to help TDNNs in their task : not only to a priori choose the kind of task that is the most likely to be successfull, but perhaps also to initialise the weights of the network using human expertise on the problem to be solved. This last point is a challenging topic for further research.

Conversely, TDNNs can learn automatically from a set of typical examples (like other neural networks); but because their architecture is speech dedicated, they certainly represent a new tool for phoneticians' investigations.

REFERENCES

[1] WAIBEL et al : Phoneme recognition using Time-Delay Neural Networks. *IEEE-ASSP, vol 37, n° 3, 1989.*
[2] HAFFNER et al : Fast back-propagation learning methods for large phonemic neural nets. *Eurospeech 89.*
[3] BIMBOT : Phonetic features extraction using Time-Delay Neural Networks. *ICSLP 90.*
[4] DEVILLERS et al : Reconnaissance monolocuteur des phonèmes du francais au moyen de réseaux à masques temporels. *XVIIIèmes JEP, 1990.*
[5] JAKOBSON et al : Preliminaries to speech analysis. *M.I.T. Press, 1951.*
[6] MALMBERG : Structural linguistics and human communication. *Springer, 1967.*
[7] ROSSI : Les traits acoustiques. *La Linguistique, vol 13, fasc 1, 1977.*
[8] CHOMSKY et al : The sound pattern of english. *Harper & Row, 1968.*
[9] COMBESCURE : 20 listes de 10 phrases phoné-tiquement équilibrées. *Rev. d'Acoustique, n° 56, 1981.*
[10] BIMBOT et al : TDNNs for phonetic features extraction : a visual exploration. *ICASSP 91.*

phonetic feature	whole corpus	vowels only	consonants only
I vowel vs non-vowel	**95.8 %** (96.2 %)	-	-
II vocalic vs non-vocalic	**96.5 %** (97.0 %)	-	-
III voiced vs unvoiced	**98.9 %** (99.4 %)	-	**98.8 %** (100 %)
IV sonant vs non-sonant	**96.9 %** (98.4 %)	-	-
V nasal vs non-nasal	**97.7 %** (99.5 %)	**97.7 %** (99.6 %)	**98.2 %** (99.7 %)
VI grave vs acute	**90.6 %** (96.5 %)	**95.7 %** (99.8 %)	-
VII extreme vs central	**84.4 %** (92.1 %)	**87.5 %** (96.6 %)	-
VIII compact vs diffuse	**91.7 %** (96.0 %)	**94.0 %** (97.8 %)	-
IX rounded vs unrounded	**94.8 %** (97.5 %)	**93.2 %** (94.5 %)	-
X bemol vs non-bemol.	**90.1 %** (98.8 %)	-	-
XI delayed vs non-delayed	**97.1 %** (98.7 %)	-	**89.4 %** (99.1 %)
XII discontinuous vs cont.	**97.9 %** (99.1 %)	-	**93.9 %** (97.6 %)
XIII fricative. vs non-fric.	**97.3 %** (99.9 %)	-	**95.1 %** (99.9 %)
Average for 6 random features	**85.3 %** (92.1 %)	-	-

Table 1 : Scores for feature extraction on the 3 corpuses : scores on test set, (on training set).

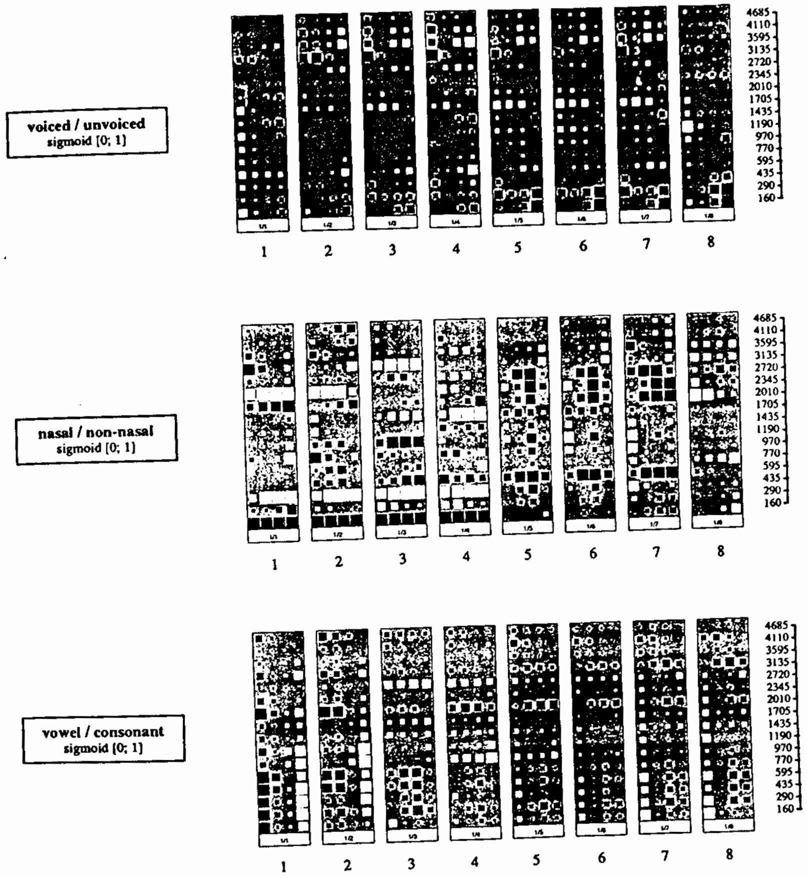

Figure 2 : Visual display of weight matrices (masks) in TDNNs for phonetic features extraction. (top : voiced / unvoiced, center : nasal / non-nasal, bottom : vowel / consonant).

The side of each square is proportional to the magnitude of the weight. Black corresponds to positive values, white to negative. Only the first layer of weights is represented, as 8 masks (4 x 16). Time is horizontal, frequency vertical (Mel-scale).

[+/- voiced] : All 8 masks are quite similar to each other, which evidences that the network is over-dimensioned for the task. The presence of energy in the first 2 bands (160 Hz and 290 Hz) is clearly used as a cue for identifying voiced patterns, especially for masks 5, 6 and 7. But a temporal decrease in high frequencies (3135 Hz and 4685 Hz bands), jointly with an increase in low frequencies is also an indication of [+ voiced] (masks 2, 3, 4, 7 and perhaps 5 and 6). This is certainly owed to phonetic combinations such as unvoiced plosive / fricative + voiced sound, for which the sudden change of the spectral tilt evidences the beginning of voicing.

[+/- nasal] : Masks show here again some redundancy. Masks 1, 2, 3 and 4 underline the significant role played by the joint presence of low frequencies in the 160 Hz band (all nasals are voiced...) and the absence (white weights) of "medium-low" frequencies (around 435 Hz). This is fully consistent with the observations concerning spectral zeros around the medium-low frequencies, for nasal sounds. These 4 masks differ mainly by the location of an other spectral hollow (2345 Hz for masks 1 and 2, 3135 Hz for mask 3, 1705 Hz for mask 4). Mask 5, 6 and 7 are very similar with one another. They search for joint energy around 595 Hz and between 2010 and 2720 Hz. This may correspond specifically to nasal vowels that usually have a first formant in [500 Hz - 700 Hz] and a high third formant in [2300 Hz - 2800 Hz]. The role of mask 8 is not clear yet (no energy around 770 Hz nor 2010 Hz).

[+/- vowel] : Here, masks 1 to 8 are ordered according to a noticeable progression : from left to right, a pattern of important energy somewhere between 290 Hz and 770 Hz (region covering roughly all possible first formants) can be approximately retrieved in all masks, with a different time-shift. Note that the lowest band is not really used, because it is ambiguous with voicing. Note also that an absence of energy around 2720 Hz (just beyond the theoretical maximum second formant for [i]), and the presence of energy around 3135 Hz (region for several third formants) are both considered as in favour of [+ vowel].