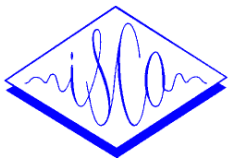

First International Conference on
Spoken Language Processing (ICSLP 90)
Kobe, Japan
November 18-22, 1990

Automatic Segmentation and Alignment of Continuous Speech Based on Temporal Decomposition Model

H.D. WANG, G. BAILLY, D. TUFFELLI

Institut de la Communication Parlée INPG / ENSERG - Université Stendhal
Unité Associée au CNRS n°368; 46, Av. Félix Viallet. 38031 Grenoble CEDEX. FRANCE

ABSTRACT

The speaker independence and the context modelling are the key problems in automatic segmentation and alignment of continuous speech, which are connected with the segmental concept of speech. In this paper, a new approach is presented: a robust speaker-independent algorithm for this task. It aligns a phonetic transcription with a phoneme nucleus detector using the temporal decomposition (TD) paradigm. The algorithm performs this task in 3 stages: a) Predetection of phoneme nuclei centers candidates using an adaptive detection window; b) Time-alignment of the corresponding phonetic transcription using a TD model based Dynamic Time Warping (TD-DTW) procedure; c) Adjustment of these output nuclei centers and phoneme boundaries detection based also on the TD model. A new temporal decomposition technique was developed also. This algorithm has been trained using 200 sentences pronounced by one speaker and tested using 50 sentences pronounced by 7 speakers. On the test corpus, 86% of the phonemes nuclei centers candidates fall into one manual segment alone. 94% of the final nuclei centers match the manual segmentation.

I. INTRODUCTION

In the speech research domain, a well-segmented and aligned acoustic database is very useful for training and evaluating a speech recognition system or for developing the rules in a speech synthesis system. Although this work can be completed manually, it is very time consuming and lacks in the consistency and reproducibility of results. Recent work [4] has tried to speed up the manual operation for broad labeling tasks. However, the consistency and reproducibility of results still calls for the development of an automatic segmentation and alignment system.

The key questions for obtaining a good performance in such a system are its independence of speakers and freedom from the influence of immediate context (effects of coarticulation) on the acoustic realization of the phonemes. Over the past decade, many automatic segmentation and alignment algorithms have been proposed. Several of these algorithms intended to align the speech signal with a reference pattern, which can be a manually labeled natural utterance[6][7], a concatenated template[3], or a synthesized utterance[8], using dynamic time warping procedures. These methods try to match the entire parametric representation of phonetic segments and thus are heavily influenced by coarticulation effects, and also depend on reference speaker. Another approach classifies the speech into broad phonetic classes before applying dynamic time alignment [9][10]. Although broad phonetic classes are robust and more or less independent of speakers, the important point is that these methods require much acoustic-phonetic knowledge to avoid the influence of context.

In fact, these problems are connected with the segmental concept of the speech. In the approaches discussed above, speech is seen as a sequential path through phonetic automata. From this point of view, the influence of coarticulation is very difficult to model. Our approach is based on a phonetic model: A Temporal Decomposition (TD) model [1][2] in which the realizations of the phonemes are seen as overlapping Phonetic Emergence Functions (PEFs): our approach consists in the alignment of the phonetic transcription with an optimal set of Phoneme Nuclei Centers (PNCs or targets in [1]) evidenced by the TD model. In this way, the influence of coarticulation can thus be modeled. For the problem of speaker independence, crude and robust parameters like energy, zero-crossings and duration functions are used. Our approach performs the task in three steps: Firstly, an algorithm for the predetection of PNCs is activated to propose PNC candidates using an adaptive detection window. Then, the output PNC candidates are aligned with the known corresponding phonetic transcription by a TD model based Dynamic Time Warping (TD-DTW) procedure. In this TD-DTW procedure, the coarticulation effect is modeled by the TD model to form a transition model between adjacent PNCs which serves as the transition cost, thus enabling insertion and elimination of the PNCs candidates, while the crude parameters cited above served to form the local matching cost. For accomplishing this kind of TD model for given targets (PNCs), a new TD technique has been developed. After this time alignment, the flow of realized phonemes is described by the overlapping PEFs of the TD model, and the time-aligned crude PNCs are adjusted to the center of gravity of each corresponding PEF. The boundaries of the phonemes are produced in the sense of the times of equal adjacent PEFs corresponding to final PNCs.

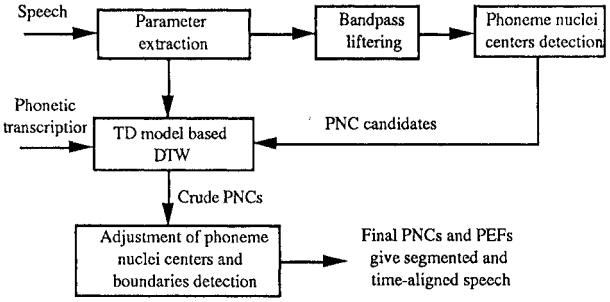

Figure 1: System structure overview

II. SYSTEM DESCRIPTION

The structure of the system is shown in Figure 1. In parameter extraction, the speech signal is first pre-emphasized with the filter $1 - 0.95\,z^{-1}$, then 14 LPC Cepstral coefficients, short term energy and short term zero crossings rate are calculated once every 5 ms in a Hamming analysis window of 20 ms. Subsequently, for effecting the first step, a bandpass liftering of the type $w(k) = 1 + \frac{L}{2}\sin\left(\frac{\pi k}{L}\right)$ for $k = 1, 2, ..., L$, where $L = 14$, is applied to these LPC Cepstrum coefficients to enforce the effectiveness and robustness of the detector [11]. Also the short term energy function and the zero-crossings function are taken respectively as following for preserving their dynamic ranges:

$$E = \sqrt[4]{\sum_{i=1}^{N} s^2(i)} \quad \text{and} \quad Z = \sqrt[4]{\sum_{i=1}^{N} \frac{1}{2N} |Sgn[s(i)] - Sgn[s(i-1)]|},$$

where

$$Sgn[s] = \begin{cases} 1 & \text{if } s \geq 0 \\ -1 & \text{if } s < 0 \end{cases},$$

and s(i) is the signal sample in the analysis window of length N samples or 20 ms. These two parameters are normalized as follows to eliminate the level influences:

$$P' = G\frac{P - P_{min}}{P_{max} - P_{min}}$$

where P_{min}, P_{max} are respectively the minimum and the maximum values of the energy or zero-crossings function in all of the speech signal, and G is the absolute gain of the parameters tuned as optimum in the PNCs detector, in the TD-DTW procedure, and in the adjustment procedure separately. The duration parameters calculated by the TD model are normalized by estimations of syllabic rate of the speech sentence.

PREDETECTION OF PHONEME NUCLEI CENTERS

The purpose of the PNCs predetection is to propose PNC candidates for the second stage of processing. It does not use any information about the corresponding phonetic transcription. This algorithm is inspired by the recent work of Van Hemert[3]. Instead of detecting transient boundaries, the center of gravity method is used to detect the PNCs in our system. Certain modifications have been made in consideration of its performance. These modifications concern the adaptivity of the threshold for the detection window (and to the immediate context), the decision criterion considering the context and detection sensibility [12]. This method is based on the following principles: two spectral frames belong to the same steady-state segment of the speech signal when they are similar to each other to some extent in certain measures. This similarity measure is defined as the Euclidian distance.

Based on this principle, detection is done using a moving window of 125 ms, i.e. with each spectral frame representing 5 ms of the speech signal, the similarity distances between the current frame and the 12 neighbours on both sides are measured. The distances curve is shown in Figure 2 (a). It is evident that this distance curve has the minimum 0 at the detection window center i.

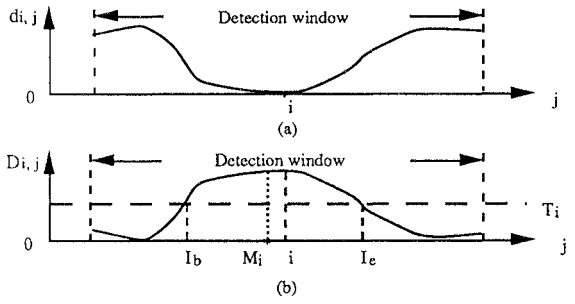

Figure 2. (a), Distances between the current frame i and the jth frame in the detection window. (b), The distances inversely transformed. A threshold T_i is taken as the half of the maximum. The frames between I_b and I_e are supposed belong to the same steady-state segment. The center of gravity M_i of this segment is indicated also.

For calculating the center of gravity, an inverse transform is applied as $D_{i,j} = Max(d_{i,j}) - d_{i,j}$, in which $Max(d_{i,j})$ is the maximum of the distances curve in the current window. After this inversion, a threshold T_i is taken as half of the maximum $Max(d_{i,j})$. Because the maximum distance value $Max(d_{i,j})$ is dependent on the context surrounding the current frame i, when the frames in the detection window are more similar ($Max(d_{i,j})$ is small), a larger number of frames (from I_b to I_e) is assumed for the same steady-state segment, and vice versa. Thus, this threshold T_i is adaptive to the detection window (context sensitive). The center of gravity of this supposed steady-state segment is calculated as:

$$M_i = \frac{\sum_{j=I_b}^{I_e} j\, D_{i,j}}{\sum_{j=I_b}^{I_e} D_{i,j}}$$

The detection function F(i) of the current detection window is thus defined as the difference between the center of gravity of the supposed steady-state segment and the current detection window center: $F(i) = M_i - i$. An example of this function is shown in Figure 3. This function can be interpreted as follows: a decrement in the detection function value indicates that the time evolution is approaching a steady-state segment and an increment in value indicates that the time evolution is leaving this steady-state segment for the next steady-state segment. The detection of the PNCs is thus the detection of the centers of the decrement part of the detection function.

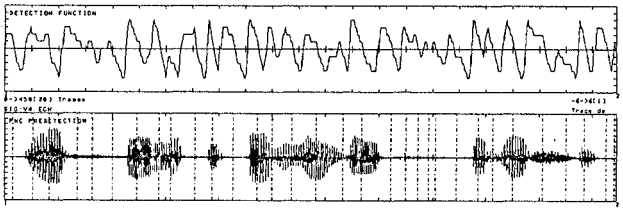

Figure 3. Example of the detection function F(i) on the sentence "La faux ne coupait rien d'autre que des choux"

TIME-ALIGNMENT BASED ON THE TD MODEL

The output PNC candidates from the PNC detector are aligned with their corresponding phonetic transcription using a TD-DTW procedure. The classic DTW algorithm, developed originally for isolated word recognition [5], has the following difficulties in application to the task of temporal alignment of these PNC candidates with the corresponding phonetic transcription:

- The transition costs (or weight coefficients in the notation of [5]) on the local path are determined *a priori* as constant under the limitation of the independence of the normalization factor upon the alignment function. Such weight coefficients can not dynamically characterize the transitions between PNCs because they have not any relation with the characteristics of the speech.
- In cases that there is a missing PNC (two or more labels correspond to one PNC candidate) or a extra PNC candidate (two or more PNC candidates correspond to one label), classic DTW has no criterion for finding the insert position, or which PNC candidate should be deleted. This requires a post-processing after the alignment.

These facts can be remedied by using the TD model.

A. A new technique for Temporal Decomposition:

The technique of the TD model for speech was first introduced by ATAL [1] for the speech coding problem. This technique has been extended to model coarticulation effects [2][13]. The main idea of the TD model is that a sequence of spectral representation of the speech signal can be modeled by a set of distinct targets and associated temporal interpolation functions (which are called here the Phonetic Emergence Functions, or PEFs). Two sorts of technique exist for this model: one is the technique proposed by ATAL [1] and its variants [13][14] which were intended to obtain the model targets and PEFs by supposing the compactness of the PEFs at certain time periods. The other one consists of the technique [2] which intends to model the PEFs with a set of given targets, but without any compactness constraint.

The new technique presented here gives an analytical solution with a fixed number of targets and also without any compactness constraint. The algorithm minimizes the reconstruction error (least squares error) between the time-frequency representation of the speech signal and the above model. In fact, the speech signal is represented as a series of spectral parameters: $P_1, P_2, ..., P_i, ..., P_N$, where P_i is a **M** dimension vector at time **i**. By supposing a set of targets: $T = (T_1, T_2, ..., T_k, ..., T_D)$, where T_k is also an **M** dimension vector. The spectrum at time **i** can be modeled as a linear combination of these targets and a set of the associated PEFs,which is a **D** dimension vector:

$$\hat{P}_i = TF_i. \qquad (1)$$

where $\hat{P}_i$ is the approximation of P_i produced by the model. This equation can be solved by introducing the criterion of least square error between the real value P_i and its model representation $\hat{P}_i$, i.e. by minimizing the scalar function $\Phi(F_i) = (TF_i - P_i)^t (TF_i - P_i)$.

By supposing the derivative of this function $\frac{d\Phi}{dF_i} = 2T^t(TF_i - P_i)$ is zero, the following equation is obtained:

$$T^tTF_i = T^tP_i \qquad (2)$$

The equation (2) is the same as that which Atal obtained [1], but here it is demonstrated as the result of minimizing the least square error between the TD model and the real speech signal. Because the parameters of the targets are known (T^tT is known), the problem is to determine whether the square matrix T^tT is invertible or not. In equation (2), T is a **M** × **D** matrix, and T^tT is a **D** × **D** matrix. If the rank of T^tT is **D**, which is the case with **D** < **M** (number of targets is inferior to the parameter dimension), the equation (2) can be inverted to yield

$$F_i = (T^tT)^{-1}T^tP_i, \qquad (3)$$

which is the unique solution. If the rank of T^tT is inferior to **D**, equation (2) has not only one solution, or has not solution at all. In this case, the problem is to find a solution, or to find a set of temporal emergence function F_i which gives the quasi minimization of the least square error. This can be done as follows: In the equation (1), $\hat{P}_i$ is the approximation of P_i, thus this equation can be rewritten as

$$P_i = TF_i + E_i, \qquad (4)$$

where E_i is the approximation error which is a **M** dimension vector. This implies the following equation

$$AX_i = P_i, \qquad (5)$$

where A = [T , I] with I a unit matrix, and $X_i = \begin{bmatrix} F_i \\ E_i \end{bmatrix}$. The rank of matrix A is **M**, and the equation (5) has infinite number of solutions.Within these solutions, one solution can be obtained by minimizing the function $\Psi(X_i) = X_i^tX_i$ under the condition $AX_i = P_i$:

$$X_i = A^t(AA^t)^{-1}P_i. \qquad (6)$$

The minimization of the function $\Psi(X_i) = X_i^tX_i = F_i^tF_i + E_i^tE_i$ (where the vector E_i is seen as independent of the vector F_i) gives an approximated solution of equation (2). Equation (6) can be simplified to obtain

$$F_i = (T^tT + I)^{-1}T^tP_i \qquad (7)$$

and

$$E_i = (TT^t + I)^{-1}P_i.$$

The equation (7) is an approximation of the equation (3). The decomposition procedure is implemented as follows: *examine T^tT, if it is invertible then apply equation (3), if not then apply equation (7).* In fact, experimentally, if the number of targets (**D**) is inferior to the dimension of the parameters (**M**), T^tT is always invertible.

B. TD-DTW algorithm:

The TD-DTW procedure takes the local path diagram which corresponds to the following recursive relation:

$$G(i, j) = \min \left\{ \begin{array}{l} G(i-1,\ j)+d_e(i, j)+w_e(i-1\rightarrow i) \\ G(i-1, j-1)+d_a(i, j) \\ G(i-1, j-2)+d_i(i, j)+w_i(i-1\downarrow i)+p_i(i-1\downarrow i, j-1) \end{array} \right\} + p(i, j),$$

where i and j are respectively the PNC candidate's number and the label number in the phonetic transcription; w_e, w_i are respectively the transition costs for the elimination of the PNC candidate i-1 and for insertion of a PNC between the PNC candidates i-1 and i; p, p_i are respectively the local matching costs for the normal alignment of PNC candidate i with the label j and for the insertion matching of the inserted PNC between i-1 and i with the label j-1, which are determined by constructing histograms of the energy and zero-crossings functions for each phoneme; and d_e, d_a, d_i are respectively the duration parameters of the PNC candidate i in cases of elimination, normal alignment and insertion. The determination of the coefficients w_e, w_i, the insert position for p_i, and d_e, d_a, d_i are based on the TD model.

Transition model

The transition costs for the transition model based on the TD model are drawn from the following observations [12] that, in modeling a short segment of speech of two or three PNCs (targets), modeling error is greater when there are less PNCs. On the other hand, modeling error when there is a lack of a PNC is much greater than when there is the just required number of PNCs, and modeling error when there is the just required number of PNCs is a little greater than when there is a redundant PNC because the missing characteristics of the lacking PNC are more critical than the superfluous characteristics of the redundant PNC for modeling this short segment of speech.

Based on this observation, the transition model is developed as follows (Figure 4): Consider three PNCs (targets) in a short speech segment, C_1, C_2, C_3, for observing the necessity of the presence of C_2 for modeling this segment by the TD model, we have two hypotheses:

- H_1: The PNC C_2 is necessary for modeling the speech segment.
- H_2: The PNC C_2 is not necessary for this modeling.

We then apply the TD model respectively to these two hypotheses, and we can obtain respectively two TD modeling errors E_1 and E_2. We define the test of the transition model by the ratio of these errors: $R = \frac{E_1}{E_2}$.

The ratio R is generally inferior to 1. When H_1 is true, R is very small (E_2 is very superior to E_1); when H_2 is true, R is only a little inferior to 1 (E_2 is a little superior to E_1). This variation of the ratio R therefore depends on the context between the three PNCs C_1, C_2, C_3, and can serve as a dynamic transition cost for the local path.

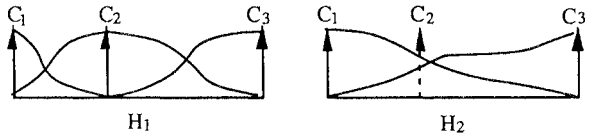

Figure 4. The Transition model.

Determination of w_e, w_i and insert position: A partial backtracking technique

This is done by applying the transition model respectively to the elimination arc (Path 1) and the insertion arc (Path 2, see Figure 5). Because in the local path we permit elimination (certain PNC candidates belong to one PNC), thus before applying the transition model we have to effect for each iteration (i,j) a partial backtracking to determine the correct PNC (noted as Key PNC) to which the PNC candidates to be eliminated belong, and also we have to determine the position of the inserted PNC. The partial backtracking is realized by an internal Dynamic Programming using the minimum TD modelling error criterion [12]. The determination of the position of the inserted PNC between two PNC candidates is realized as the position where the TD modeling error between these two PNC candidates is maximum. For example (Figure 5), for modeling the elimination arc we will take the PNC candidate i as the target C_3, the key PNC for the candidates from i-m to i-1 as C_2, and the key PNC for the candidates from i-h to i-m-1, or the PNC inserted between the candidates i-m-1 and i-m, as C_1; for the insertion arc, we take the PNC candidate i as the target C_3, the key PNC for the candidates from i-n to i-1 as C_1, and the PNC to be inserted between the PNCs candidates i-1 and i as C_2.

After applying the transition model like this, the test ratio of the transition model for each arc is taken as the transition cost for the corresponding arc, i.e. w_e and w_i.

Determination of the duration parameters d_e, d_a, d_i

These phonemic duration parameters are obtained simply by taking the duration between the times when the PEF for a PNC candidate is equal to those for its neighbours in the TD modelling. The matching of these duration parameters with the phonetic transcription labels are realized by means of histograms for each phoneme.

Simplified implementation of TD-DTW

The implementation of the TD-DTW algorithm described above is very time consuming. We can simplify this partial backtracking technique by

taking an approximated PNC as the key PNC to which the PNC candidates to be eliminated belong instead of using the internal dynamic programming determination. For example, for applying the transition model to the elimination arc, we take the three targets as follows: C_1 is taken as the PNC candidate i-2, C_2 as i-1, and C_3 as i; for applying the transition model to the insertion arc, we take the three targets as: C_1 as i-1, C_2 as the spectral vector at the position where the TD modelling error between PNCs i-1 and i is maximum, and C_3 as i. This simplification will slightly degrade the performance of TD-DTW, but reduce significantly the calculating time.

This alignment requires that the phonetic transcription corresponds exactly to the acoustic realization, including the interword pauses. Figure 6 shows an example of the results after each processing stage.

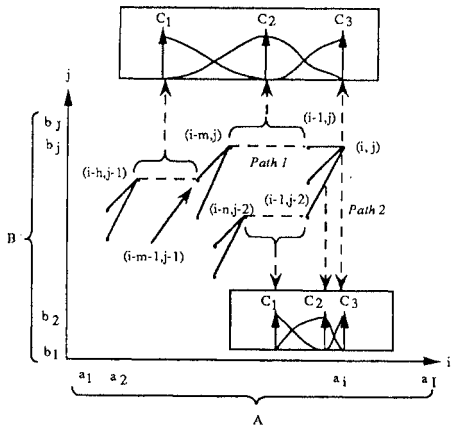

Figure 5: The partial backtracking technique

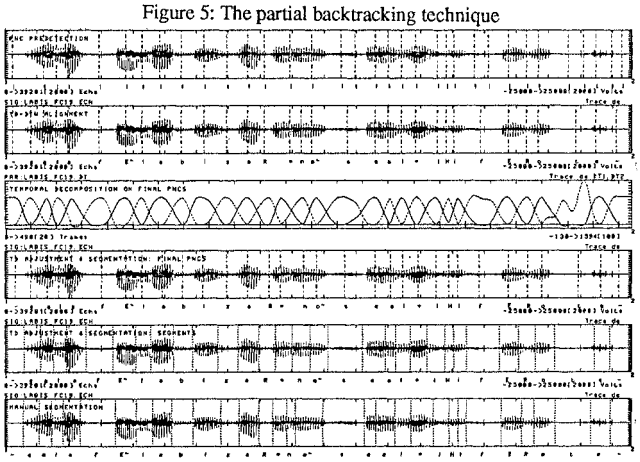

Figure 6. Results of PNCs after different steps on the sentence "Et à la fin la bise a renoncé à le lui faire oter": (a). PNCs predetection. (b). TD-DTW alignment. (c). the PEFs of the PNCs after the adjustment. (d). adjustment of TD-DTW output PNCs. (e). the phoneme boundaries after the adjustment. (f). results of manual segmentation.

ADJUSTMENT OF THE TIME-ALIGNED PHONEME NUCLEI CENTERS AND BOUNDARIES DETECTION

The output time-aligned PNCs are not exactly situated at the centers of the realization of the phonemes because of the insertion and the elimination of PNC candidates, thus requiring an adjustment. This can be done by replacing them at the center of gravity of the corresponding PEFs. The calculation of the center of the gravity of the PEFs uses the same formulas as in the PNC detector except that the threshold here is taken as 80 percent of the maximum value of each PEF. An example is shown in Figure 6 (d). For example in Figure 6 (b), the PNCs "*" in context "R * n o~" and "i" in context "H i f E" are not well aligned after the TD-DTW alignment procedure, by this adjustment (shown in Figure 6 (d)), these targets are changed to the correct positions of the PNCs. Because the acoustic realizations of nonsonorant plosives as "p", "t", "k", and of silence in speech signal are poorly modeled by the LPC Cepstral coefficients, the corresponding PEFs are heavily influenced by noise. We adopt not to reposition these targets.

When this adjustment of the PNCs has been done, generally, the realization of the phonemes are well modeled by the PEFs overlapping in time with their neighbours [2] (Figure 6 (c)). Thus the time that the adjacent PEFs equal to each other, is taken as the boundary between the two corresponding phonemes. This result is also shown in Figure 6 (e).

III. EVALUATION

The previously described algorithm has been trained using 200 sentences by one speaker and tested using 50 sentences pronounced by 7 other speakers. On the test corpus, 86% of the PNC candidates fall into one manual segment alone. This result for the PNC detector is obtained by tuning the detection window to an optimum length, thus the extra-detection and miss-detection errors are balanced (both 7%). Considering that the miss-detection error is more grave, we tune this parameter in order to obtain the best performance for the complete system, and this results a 2% extra-detection error and a 21% miss-detection error. From this resulting PNC detector, 94% of the final PNCs match the manual segmentation. In comparison with the classic DTW algorithm (we have replaced the dynamic transition costs introduced by the transition model based on the TD model by the weight coefficients determined *a priori* as the constant 1), the TD-DTW algorithm raises significantly the performance (more than 5%).

IV. CONCLUSION

In this paper, a new approach of automatic segmentation and alignment of the speech signal with its corresponding phonetic transcription is described. This approach performs the segmentation and time alignment task by working with the phoneme nuclei centers, and using the temporal decomposition model to describe the phonemes. The crude parameters of energy and zero-crossings enable this system complete the task with independence of speakers. Although this system requires that the phonetic transcription corresponds exactly to the realization of the phonemes, this system can be used to segment and label large databases[15].

REFERENCES

[1] ATAL B. S., "Efficient coding of LPC parameters by temporal decomposition.", Proc. ICASSP, 1983, pp.13-16.

[2] BAILLY G., MARTEAU P. F., ABRY C., "A new algorithm for temporal decomposition of speech. Application to a numerical model of coarticulation." , Proc. ICASSP, 1989, pp.508-511.

[3] VAN HERMERT J. P. "Automatic segmentation of speech into diphones.", Philips Technical Review, Vol. 43(9) September 1987, pp.233-242.

[4] TUFFELLI D., WANG H. D. "TELS: A Speech Time-Expansion Labelling System.", In this Proceedings.

[5] SAKOE H., CHIBA S., "Dynamic programming optimization for spoken word recognition." IEEE Trans. Acoustic. Speech. Signal Processing, Vol. ASSP-26, Feb. 1978, pp.43-49.

[6] CHAMBERLAIN R. M., BRIDLE J. S., "Zip: A dynamic algorithm for time-aligning two indefinitely long sequences.", Proc. ICASSP, 1983, pp.816.

[7] HOHNE H. D. et al., "On temporal alignment of sentences of natural and synthetic speech." IEEE Trans. Acoustic. Speech. Signal Processing, Vol. ASSP-31, August 1983, pp.807-813.

[8] LENNIG M., "Automatic alignment of natural speech with a corresponding transcription.", Speech Communication, Vol. 2(3), 1983, pp.190-192.

[9] WAGNER M., "Automatic labelling of continuous speech with a given phonetic transcription using dynamic programming algorithms.", Proc. ICASSP, 1981, pp.1156-1159.

[10] LEUNG H. C., ZUE V. W., "A procedure for automatic alignment of phonetic transcriptions with continuous speech.", Proc. ICASSP, 1984, pp.2.7.1.

[11] JUANG B. H., RABINER L. R., WILPON J. G., "On the use of bandpass liftering in speech recognition.", IEEE Trans. Acoustic. Speech. Signal Processing, Vol. ASSP-35, July 1987, pp.947-953.

[12] WANG H. D., "Segmentation et Etiquetage Automatique de Base de Données des Sons du Français", Ph.D. Thesis, Institut National Polytechnique, Grenoble France, 1990.

[13] AHLBOM G., BIMBOT F., CHOLLET G., "Modeling spectral speech transitions using temporal decompositions techniques.", Proc. ICASSP, 1987, pp.13-16.

[14] VAN DIJK-KAPPERS A. M. L., MARCUS S. M., "Temporal decomposition of speech.", IPO annual progress report, 22, 1987, pp.41-50.

[15] BAILLY G., BARBE T., WANG H. D., "Automatic labelling of large prosodic database: Tools, methodology and links with a text-to-speech system", Proc. of the 1st Int. Workshop on Speech Synthesis, Autrans, France, (to appear).