

Very-low-rate speech compression by indexation of polyphones.

Charles du Jeu, Maurice Charbit, Gérard Chollet

GET-ENST, CNRS-LTCl, TSI Dept, 46 rue Barrault, 75634 Paris cedex 13 - France

dujeu@tsi.enst.fr
charbit@tsi.enst.fr
chollet@tsi.enst.fr

Abstract

Speech coding by indexation has proven to lower the rate of speech compression drastically. Based on the Automatic Language Independent Speech Processing (A.L.I.S.P) approach that automatically segments the speech signal ([1]), we studied the possibility of optimising this rate as well as the quality of re-synthesised signal, by using the text information corresponding to the speech signal, and by implementing a new segmentation method. This led to the speech alignment with its phonetic transcription and the use of polyphones, to finally increase output speech quality while keeping a bitrate between 400bits/s and 600bits/s. Typically, this can be used to store recorded alpha-numeric books for blind people, or compressing recorded courses for e-learning. Cell phone applications could also be considered.

1. Introduction

The main goal of this work was on one side the implementation a new speech segmentation in the A.L.I.S.P (Automatic Language-Independent Speech Processing) system developed at E.N.S.T ([1]), and on the other side the improvement of the efficiency of the system, exploiting text information.

Indeed, a problem with the current A.L.I.S.P system is the segmentation step : Temporal Decomposition to extract coherent segments from the speech signal implied segmentation on particularly instables zones (stability will be discussed later in this paper, in section 4.1). Thus, acoustical result of the concatenation in those zones was not satisfying, so we explored another way of segmentation by cutting segments, on opposite, on stable zones. This approach, first implemented by Baverel ([2]), increased drastically the quality of output speech.

Assuming we know which text is said, either by automatically recognising it or because it is given *a priori* in specific applications, it is possible to align the speech signal with its phonetic transcription, and to use language-specific knowledge to recognise segments as "polyphones" : sequences of N phones, typically $0 < N < 3$, and $\max_N = 10$. We exploit here the work of Frédéric Bimbot on french polyphones ([3]).

After a brief overview of the main steps of the system (section 2), we will explain in detail each part of the process (section 3 to section 6). Then, the bitrate will be studied in section 7, and a comment on the quality of the new synthesized speech will be discussed. As a conclusion, we will suggest some improvements to get more robust algorithms.

2. Main principles

In a compression process based on segmental indexation, before any coding operation, a reference database of acoustic segments

should be created, that will be present in the coder and in the decoder. This implies the availability of a consequent amount of physical memory on both sides.

. First, the *training* process builds up a database (DB) computed from a large amount of data called *corpus*, by extracting acoustic segments, classifying them and storing the most relevant ones.

. Then the *coding* process segments with the same method the speech file to be coded and recognizes each segment as belonging to one class of the DB. Each segment is compared to all representants of its class in the database and the closest one is chosen. Thus, the transmission data for one segment will only consist of a reference to the representant in the DB, and some side information tracing the differences between this representant and the original segment.

. Last, the *decoding* process : assuming the decoder has the same DB on its side, the representants identified by the transmitted data are modified, concatenated and finally resynthesized to speech.

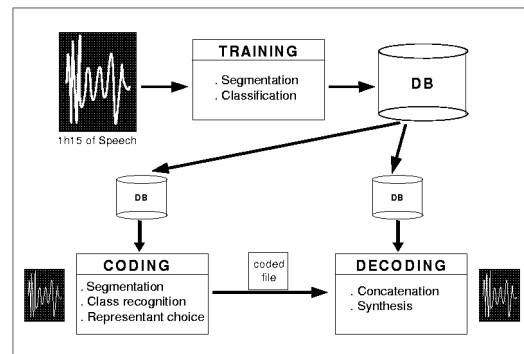


Figure 1: The main steps of the compression process: training (segmentation, classification), coding (segmentation, recognition, choice of best example) and decoding (concatenation, synthesis)

3. Parametrization of signal

In our experiments, the original speech files are sampled at 16kHz Mono with a 16bits resolution.

3.1. Harmonic Plus Noise Model (H.N.M.)

The system uses the "Harmonic plus Noise Model" for the voice signal, *i.e.* speech is represented as the sum of a harmonic part

(which is obviously the segment itself), the routine selects the second closest representant : by this way, we obtain a frequency of selection for each representant. By deleting the unused segments, the size of DB is consequently divided by two. Furthermore, keeping a fixed number of segments in each class, *e.g.* by keeping the most selected ones, can lower much more the size of the database, but the quality of output speech is spoiled very rapidly.

5. Coding

The speech signal to be coded is segmented as previously explained. Since the classification is phone-based, class is fully determined by the phones that compose each segment. Once a known class corresponds to every segment of the file to code, Dynamic Time Warping (see original Sakoe and Chiba [6]) is performed for each segment between the original one and all the representants of the class. Dynamic Time Warping consists of finding the best path through a network composed by the two sequences to compare. It gives a numerical estimation of the *distance* between two segments of different lengths, and the best path establishes a correspondance between each frame of the original and a frame of the example. The representant which has the smaller distance is selected and the dtw-path to this representant can be communicated to the decoder. But to keep the bitrate very low, **the DTW algorithm is used only on the coding side to find the closest representant**. For this reason, we also improved the algorithm by setting a new constraint on the slope of the path so that it keeps closer to the diagonal path (see also decoding section). **Fig 3** shows the paths computed by the old algorithm and by the new one. More, our new algorithm is faster, since we compute the cumulated distance at less nodes of the network.

Thus a sample code contains for each segment : the position of the first frame of the segment, its length, the class and the number of the best representant in this class, and the original prosody, that is, the energy and the pitch.

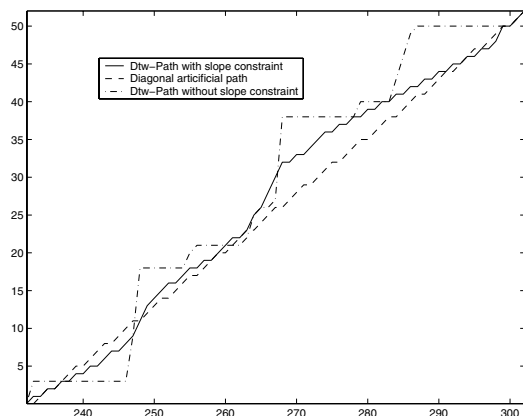


Figure 3: *Dynamic Time Warping : finding the best path through a network. We compare here the path computed by the old AL-ISP algorithm (without slope constraint), the artificial diagonal path that will be used on the decoder side, and the new path computed by our algorithm.*

If a polysound for one segment is not found in the database, that is, no occurrence of such segment was extracted during training, a routine of re-segmentation splits this segment into

different combinations of polyphones and search for the best one in the database : since most unknown segments are likely to be long polyphones, and if the training data is long enough, it is possible to find a combination of existing classes of **bi-** or **tri-** sounds in the database to recompose the same polysound. The routine automatically determines the best combination using DTW distance between original segment and every possible existing combination. Then, the DTW path allows to place a new segmentation instant inside the original segment.

6. Decoding

The decoding section is simple : for each segment, the representant is aligned close to the original thank to the diagonal dtw-path, and the correction gain factors and pitch interpolation are applied (see section 7). All segments are concatenated to form a new HNM file. Speech is then synthesized with those new HNM parameters. In this sense, the decoder program could be assimilated to a Text-to-Speech program using the prosody of an original recording to give more liveness to speech. The simplicity of the decoding stage can be useful for any broadcast mobile application processed on lightweight terminals which cannot handle heavy memory usage.

7. Bitrate

7.1. Reference to the representant

The bitrate needed for referring one segment to its class is directly linked to the number of classes. Thus, coding the identifier of the class on 11bits will be enough : the maximum number of classes is 4096, which is more than what was foreseen in section 4.4 (generally about 3500 classes).

We chose to code the reference of the representant inside a class on N bits, N a number depending on each class; N is derived *offline* from the number of representants for each class. A simple mean-value of the number of representants per class, excluding the "singletons" (classes containing only one representant), gives 13 representants per class for our corpus, which suggests coding this identifier on 4 bits. By studying precisely the content of each class and deducing the bits needed for each class (excluding the singletons, where 0 bit will be needed to code the representant), we obtain an average of 4.2 bits/segment. Taking the singletons into account gives an average of 1.9bits. However it is likely that this rate should be greater, since the frequent classes are also those containing most segments. In our experiment, coding all the test files gives an average bitrate of **6bits/segments** for this parameter. We may expect better results using coding (*e.g.* Huffman) which take into account class frequencies.

The reference to the database segment can thus be coded on $(11 + 6)bits/seg * 10seg/s = 170bits/s$.

7.2. Segment Length and Dtw-Path

The original length can be compared to the length of the representant. One bit for the sign of the difference, and 5 bits for its value are enough for this parameter. So coding the length then needs a bitrate of $60bits/seconds$. We choose **not to transmit the dtw-path**, taking only the diagonal path to align representant length on original length on the decoder side.

7.3. Coding of the prosody

The pitch and energy of each frame cannot be transmitted without quantization to the decoder, otherwise the initial bitrate goal will be overwhelmed. The principle of their coding follows : for the energy, a unique correction gain factor is transmitted per segment instead of one per frame, re-adjusting the representant to be closer to the original. For the pitch, a linear interpolation of pitch trajectory is computed.

Two gain correction terms are sent, one of the noise part and one for the harmonic part. These two parameters are sent on 7bits each, to keep enough precision.

The interpolation of pitch allows to transmit for one segment only the mean-value and the slope coefficient a given by the following relation :

$$f(n) = (a * n + b) * fr(n),$$

where $f(n)$ is the pitch of the original at frame n and $fr(n)$ the pitch of representant at frame n . The parameter b is not transmitted, because the mean-value of $f(n)$ is transmitted, and the mean-value of $fr(n)$ is available offline. Coding the mean-value on 7bits, to have enough precision, and the parameter a on 3bits gives good results.

So the definitive mean-bitrate for this coder is **470bits/s**. See table 2 for details.

Parameter	bits/segment
Name of class	11
Number of rep.	6
Length of rep.	6
Gains correctors	14
Pitch mean-value	7
Interpolation parameter a	3
Total	47

Table 2: Bitrate needed for each parameter for one segment

8. Results

For each hearing test, we present to listeners our new coded and decoded speech, the HNM analyzed and synthesized version, the old A.L.I.S.P. version, and the original Wav file. As expected, comparing results of the old A.L.I.S.P system which used Temporal Decomposition, and results of this system, the concatenation between segments sounds much better and the global hearing intelligibility is increased. Tests were made on two examples : the voice of Cathy (professionnal speaker) reading sentences of french newspaper **Le Monde** and also the voice of André Dussolier (French actor) reading Victor Hugo's book **Notre-Dame de Paris**([7]). These tests have shown the feasibility of such a program, by giving very good results with the voice of Cathy (its phonetic transcription is very close to the pronounced speech), but also its limits and the improvements to be done : indeed, the voice of André Dussolier involved problems. This voice is very dynamic in terms of frequency and energy. Most, in this case a lot of usually purely voiced frames are very noisy compared to an average french speaker. This implies a lot of errors in phonetic labellings as well as misbehaviour of HNM synthesis, which is not as transparent as expected. To improve the system, we will have to focus on generalisation and improvement of HNM implementation, and also on the robustness of phone-labelling.

9. Conclusion

The indexation of polyphones segmented on cepstrally-stable instants drastically improves the quality of resynthesized speech signal, above all in terms of intelligibility. However, it requires a robust alignment between speech and its phonetic transcription, for any speaker and any language. Since the decoder is appartened to a Text-To-Speech system, it could also be interesting to search for patterns of prosody into the sequences of polyphones on the coder side, and eventually link them to pre-defined "emotions".

Applications of this compression to recorded books could enhance their accessibility to blind people, and create a secure broadcast system on the internet : since the user needs the database of the speaker to actually listen to the book, a system in which the database is billed and sold on a single compact disc can be developped, then every coded book can be downloaded rapidly through the network, and charged free or very cheap. One can even imagine to sell a CD containing the integrality of the work of an author, or a collection of great workpieces of literature, etc. It would also be very simple to link the code with text to perform a synchronisation of the text with the reading, and for example project it on a big screen for partially-sighted persons. Broadcast on cellular phones could also be thought of, e.g. by sending voice messages on the SMS channel.

10. Acknowledgments

This work was supported by the French Ministry of Research as part of the RNRT-SYMPATEX project.

11. References

- [1] G. Chollet, J. Cernocký, et al. : *Towards ALISP: a proposal for Automatic Language Independent Speech Processing*, in Computational Models of Speech Pattern Processing, Berlin, DE, Springer Verlag, 1999, p. 375-387, ISBN 3-540-65478-X.
- [2] C. Baverel, P. Gournay, G. Chollet : *Amélioration d'un codeur de parole très bas débit par indexation d'unités de taille variable*. 18th Colloque GRETSI, Toulouse, France, Sept. 2001.
- [3] F. Bimbot : *Synthèse de la parole: des segments aux règles avec utilisation de la décomposition temporelle*. E.N.S.T Paris, Thèse de Doctorat, 1988.
- [4] I.Stylianou : *Modèles harmoniques plus bruit combinés avec des méthodes statistiques pour la transformation de la parole et du locuteur*. E.N.S.T Paris, Thèse de Doctorat, 1996.
- [5] S. Young, G. Evermann, D. Kershaw, G. Moore, J. Odell, D. Ollason, V. Valtchev, P. Woodland *The HTKBook - Revised for HTK Version 3.1 December 2001*. Cambridge University Engineering Department, 2001
- [6] H. Sakoe, S. Chiba : *Dynamic Programming Algorithm Optimization for Spoken Word Recognition*. IEEE Transactions on Acoustics, Speech and Signal Processing, Vol. ASSP-26, No.1, Feb. 1978
- [7] V.Hugo : *Notre-Dame de Paris*, 1831.
- [8] F. Yvon : *Prononcer par analogie: motivations, formalisations et évaluations*, Thèse de doctorat, ENST, Paris, May 96.