AUTOMATIC DIPHONE EXTRACTION FOR AN ITALIAN TEXT-TO-SPEECH SYNTHESIS SYSTEM

Bianca Angelini (*), *Claudia Barolo* (**), *Daniele Falavigna* (*), *Maurizio Omologo* (*) *and Stefano Sandri* (***)

(*) IRST - Istituto per la Ricerca Scientifica e Tecnologica, 38050 Povo di Trento, Italy
(**) Eikon Informatica, Via Sostegno 65/bis, 10146 Torino, Italy
(***) CSELT - Centro Studi e Laboratori Telecomunicazioni S.p.A., Via G. Reiss Romoli 274, 10148 Torino, Italy

ABSTRACT

This paper describes a system for the automatic extraction of diphone units from given speech utterances. The method is based on an automatic phonetic segmentation and on a subsequent rule-driven diphone boundary detection. The phonetic segmenter, developed at IRST, was trained and tested both in speaker independent and speaker dependent mode. A rule formalism, involving acoustic parameters, arithmetical and logical operators, was defined to express the acoustic/phonetic knowledge acquired during previous experiences on manual diphone segmentation. A specialized tool for rule parsing was designed that processes a given sequence of automatically derived phone boundaries using a corresponding sequence of predefined acoustic parameters. Several sets of rules were developed that include both general principles and specific details concerning the content of the diphone database of "Eloquens"®, the CSELT text-to-speech synthesis system for the Italian language. The accuracy was evaluated by comparing the manual and the automatic segmentations of the speech utterances of a female speaker, resulting in nearly 95% of correct boundary position, given a tolerance of 20 ms.

1. INTRODUCTION

In concatenative text-to-speech synthesis systems the acoustic dictionary represents a fundamental part, since it is strictly related to the resulting quality and naturalness. The design and collection of speech segments for the dictionary usually requires a lot of time for manual segmentation and labeling of speech databases.

In the literature, several automatic methods for unit extraction have been proposed. They often adopt techniques derived from speech recognition [1], [2]. In fact, the statistical modeling of speech has proved to be advantageous in speeding up the acoustic dictionary generation, favouring the expansion of the number and structure of speech units and encouraging the extension to different voices and languages. However, the accuracy of statistical approaches strongly depends on parameter estimation procedures, training database sizes, whereas a lot of specific and detailed acoustic/phonetic knowledge risks to be poorly represented.

This work originated with the aim of merging the IRST experience on speech segmentation algorithms [3] [4] with the CSELT acoustic/phonetic knowledge, acquired during manual extraction of speech unit inventories for an Italian text-to-speech synthesis system [5].

An automatic method for obtaining a predefined set of diphone boundaries, starting from a given set of utterances and from their corresponding phonetic transcriptions, will be described. It is realized through the two following steps:

1) automatic segmentation and labeling according to the phonetic transcription;
2) automatic diphone boundary location by application of acoustic/phonetic rules.

The performance of both the phonetic and diphonic segmenter will be discussed, by comparing manually and automatically located boundaries at different error tolerances.

2. THE PHONE SEGMENTATION AND LABELING SYSTEM

The automatic phonetic segmentation and labeling module derives from the speaker independent continuous speech recognizer developed at IRST [3] and based on Continuous Density Hidden Markov Models (CDHMMs). Each phone is modeled by a context independent HMM. The input to the segmentation system consists of a speech waveform (sampled at 16 kHz) and the corresponding phonetic transcription. The acoustic analysis is performed every 5 ms, using a 20 ms Hamming window, with preemphasis factor 0.95. Eight mel-scaled cepstral coefficients are computed from the output of a 24 triangular bandpass filterbank, together with the corresponding first and second order time derivatives; the normalized log-energy and the corresponding time derivatives are also computed for each frame.

The reference phone set is the one adopted for the development of the CSELT text-to-speech synthesis system. It includes 17 vowel-like phones (stressed, unstressed and reduced vowels, semivowels and semiconsonants), 44 consonant phones (including some allophones and geminates), two types of pause identifiers and a "schwa" symbol. For HMM training, these phones are grouped into a smaller set consisting of 36 phonetic units. Each HMM has a five state left-to-right topology (a three state topology was used only for short semiconsonants /j/ and /w/). To derive this phone subset, reduced vowels were assimilated with their unstressed counterparts, while distinct unit models were adopted both for stressed and unstressed vowels. Geminates share the same HMMs with the corresponding single consonants (with the exception of geminate /r:/, whose acoustic structure is somewhat different from non-geminate /r/).

3. PHONE SEGMENTATION PERFORMANCE

In order to obtain a robust and consistent set of phone HMMs, a suitable number of occurrences of each phone must be present in the training database. On the basis of the previous

experience [4], the minimum number of occurrences, to ensure an adequate system accuracy as well as a fair performance evaluation, is 30 for training and 5 for testing.
Phonetic alignment experiments were carried out both in speaker independent and in speaker dependent mode.
In the former case, the APASCI corpus [6] was used. This corpus consists of syntactically consistent (even if meaningless) sentences, designed to ensure a wide coverage of phonetic contexts. A subset of this corpus, consisting of 200 sentences (uttered by 50 speakers, 25 males and 25 females), was manually segmented. The phone HMM training and the system evaluation were accomplished by using 150 and 50 sentences, respectively.
The analysis of the discrepancy between the automatically determined phone boundaries and the corresponding manually segmented ones provided an average rate of 90.5% correct boundaries, given a tolerance of 20 ms.
A second database, collected at CSELT laboratories, was used for a speaker dependent evaluation with similar phonetic coverage constraints. In this case, 200 sentences (about 7500 phones) for HMM training and 50 sentences (about 1960 phones) for performance evaluation were uttered by one male (M1) and one female (F1) professional speaker. An expert phonetician manually aligned each utterance with the corresponding phonetic transcription: Fig. 1 reports on the percentage of correct automatic placement of phone boundaries, for different tolerances. The automatic alignment performance is almost independent of the speaker, and it does not change significantly, for tolerance intervals larger than 10-15 ms.

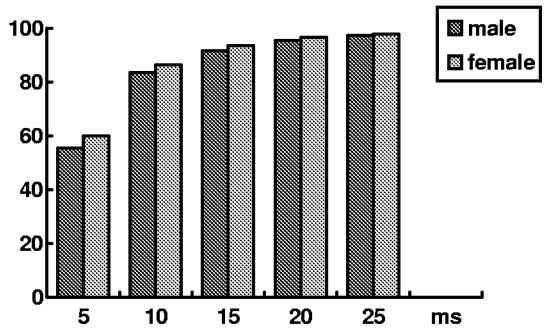

Fig. 1 - Speaker dependent phone segmentation rate

These results are not homogeneous across different classes of phone sequences, as shown in Tab. 1 for the most representative phone-to-phone transitions. For transitions including stops, fricatives and vowels, the system performs very well; performance is somewhat inferior in the cases including nasals and liquids; finally, for vowel-to-vowel transitions the lowest accuracy was observed. In the latter case, besides a possible lack of training material, the main reason of this performance loss is the intrinsic ambiguity, in boundary placement, that characterizes manual segmentation as well [7].
It is worth noting that, when using speaker dependent HMMs, the automatic alignment system provided a definite improvement: rates of 95.6% and 96.8% were obtained for M1 and F1, respectively, whereas 87.4% and 88.6% were obtained using the system in speaker independent mode. This discrepancy is due both to the different acoustic environment characteristics (between training and testing conditions) and to an intrinsic more robust modeling that is attained in speaker dependent mode. Moreover, while M1 and F1 are two professional speakers, the speaker independent system was trained using speech material uttered by naive speakers.

	5ms	10ms	15ms	20ms	N
stop-vowel	64.8	97.4	99.3	100.0	273
fricative-vowel	68.0	96.7	99.4	100.0	181
liquid-vowel	50.0	82.9	92.4	96.5	170
nasal-vowel	48.0	79.4	95.1	98.0	102
vowel-stop	73.0	93.4	97.6	98.6	211
vowel-fricative	69.1	93.2	97.5	98.8	162
vowel-liquid	66.9	83.1	91.6	97.0	166
vowel-nasal	59.6	85.1	92.5	95.7	161
vowel-vowel	24.8	48.9	69.3	81.8	137

Tab. 1 - Correct boundary placement rates across phone-to-phone transition classes in speaker dependent mode.
N indicates the number of occurrences observed for each transition class.

A further test database of 53 sentences (about 1860 phones) was considered for evaluating the robustness of the speaker dependent HMMs against their use with other voices selected among a restricted number of professional speakers: to this purpose another female voice (F2) was added to M1 and F1. Given a 20 ms tolerance, the two speaker dependent cases (M1-M1 and F1-F1) provided reference rates of 94.2% and 95.2%, respectively. Changing the test speaker caused different system behaviour: the performance reduction was small for speakers of the same sex (F2-F1: 3.1%), while somewhat greater in the other cases (M1-F1: 4.4%), in particular when the system was trained with a male reference speaker (F1-M1: 7.4%; F2-M1: 6.9%). Nevertheless, system performance was better than using the speaker independent HMMs.

4. DIPHONE SEGMENTATION

The Eloquens® Italian text-to-speech synthesis system is based on about 1100 diphone units [5]. Each of them includes transitions between a couple of half-phonemes, usually bounded around the corresponding pseudostationary parts.
A specific corpus was designed, with the following controlled characteristics for each diphone typology:

- phonetic and syllabic patterns of the word;
- diphone position inside the word;
- distance from lexical stress (for unstressed diphones);
- distance from lexical and syntactic boundaries.

In a classical implementation with a male voice (M1), isolated nonsense words were used [5]; in a recent version with a female voice (F1), lexical words containing three and four syllables were selected and embedded in syntax-controlled meaningful sentences. The first content word, following an article or a preposition, was usually preferred for its uniformity in terms of spectral shapes, articulation, pitch and intensity. The candidate diphones are specific parts of speech to be extracted from stressed or unstressed syllables of

such words according to given constraints. Diphone segmentation was carried out manually from a corpus of about 700 sentences, taking into account waveform and spectrogram observations, phone segmentations, energy contours, formant trajectories and timing details. Diphone labeling consists of two boundary markers and one transition marker, that separates the two half parts of the phonemes; pitch markers were assigned to voiced portions. Each particular phone or class of phones was subjected to proper acoustic criteria for diphone boundary location, also incorporating specific details of the text-to-speech synthesis design. Spectral stability and local maxima of signal energy as well as duration constraints for vowels, half closure interval for single stops, pre-noise attack for unvoiced affricates, central time positions for fricatives, local energy minima for liquids represent some examples of the general criteria involved.

5. ACOUSTIC/PHONETIC RULES

A rule-based mechanism has been designed with the purpose of transferring the acoustic/phonetic knowledge described above into an automatic segmentation system easily applicable to enlarged diphone sets and to different voices. Each rule can refer either to a specific diphone (e.g. [p_r]) or to diphone classes (e.g. [stop_liquid]) or to a mixture of them (e.g. [stop_r]), through a declaration mechanism, where single phone symbols can be grouped into phone classes.
A generic rule is expressed in terms of "*acoustic parameters*", "*conditions*" and "*operators*".
The set of "*acoustic parameters*", evaluated over 20 ms Hamming windows, at 5 ms steps, includes:

- the signal energy (E);
- an 8-th order Spectral Variation Function (SVF), as defined in [3];
- the relative time position (D) inside a phone.

All these parameters are normalized in order to assume values in the interval [0,1]. The signal energy serves as indicator of local maxima and minima and is used in contexts where abrupt intensity changes are essential to locate a boundary (e.g. /r/ and /r:/). The SVF parameter allows the detection of fast spectral changes as well as of slow spectral transitions, being correlated with formant movements: a low SVF value usually indicates nearly steady spectral characteristics. Finally, the D parameter, expressed as a percentage of the phone duration, identifies a specific time position inside a phone.
A "*condition*" allows to determine either a specific frame or an interval of values for a given parameter, where a rule can be applied. The set of "*conditions*" can be represented by the following symbols:

- "==" specification of a value for a given parameter;
- "<X1,X2>" specification of the range of values [X1,X2] for a given parameter.

The "*operators*" are symbols used to combine logically the above mentioned "*conditions*" on the "*acoustic parameters*". The set of available "*operators*" is:

- "&" (logical AND);
- "| " (logical OR).

As an example, a rule can be expressed as follows:

#defrule left_rule [diphone_type] right_rule

where "left_rule" and "right_rule" may be:

P1 == V1 operator P2 <X1,X2>

The first *condition* (P1 == V1) searches for the frame where the *parameter* P1 assumes the closest value to V1. The second *condition* (P2 <X1,X2>) verifies if in a given frame (e.g. the one resulting from the previous *condition*) the value of the *parameter* P2 is inside the interval [X1,X2]; finally, *operator* links logically the two *conditions*, so as the rule is applied only if the logical constraints imposed by *operator* are satisfied.
The "diphone_type" specifies the left and right phones (or phone classes) to which the "left_rule" and "right_rule" are independently applied, respectively.
For example, the following segmentation rule:

SVF==0.0 & E<0.8,1.0> [vowel r] E==0.0 & D<0.3,0.7>

applies to all vowels followed by /r/; the left boundary is placed at the frame where the SVF parameter takes its minimum value, only if the energy parameter is within the range 80-100%; the right boundary is positioned at the minimum energy frame, only if it is around the central part of /r/, in the range 30-70% of the phone duration. If all the conditions are satisfied the rule is applied, otherwise it is skipped.
Rule interpretation is accomplished by a left-to-right rule parser, connected with the data structure of the phonetic segmenter. The file of segmentation rules is a plain ASCII text, formed by two sections:

1) definition of the phonetic groups;
2) list of the diphone dependent rules.

Rules must be ordered, as the list is sequentially processed starting from the most general ones and ending with the most detailed ones. Sets of rules can be easily augmented with more and more detailed conditions and values. For a given diphone type, the last applied rule is retained.

6. EXPERIMENTAL RESULTS

Various sets of phonetic groups and segmentation rules were defined and experimented on the female diphone inventory, according to the following priority ordering:

- D-rules, with different assignment values to acoustically homogeneous phonetic groups (vowels, stops, fricatives, affricates, nasals, liquids, ...);
- SVF-rules plus D-range for vowels, liquids, nasals and voiced fricatives;
- E-rules plus D-range for vowels, nasals, /r/ and /r:/;
- SVF-rules plus E-range for specific liquid allophones;
- D-rules plus E-range for particular contexts of /r/;
- D, E, SVF rule extension with increasing number of assignment and range values and more detailed phonetic groups (simple/geminate, voiced/unvoiced consonants, stressed, unstressed and reduced vowels, semivowels, semiconsonants, "schwa", ...).

The complete set of rules consists of 32 phonetic group definitions and 167 overall rules. The left and right rule fields are not always simmetric because of some phonotactic and allophonic restrictions.
The rule system was tested by comparing manual and automatic diphone boundary segmentations for the female voice. Fig 2 reports on system performance evaluated on left and right boundaries separately. As shown in the Figure, there is a slight difference between the two cases, mainly due to the independent left/right rule application mentioned above.

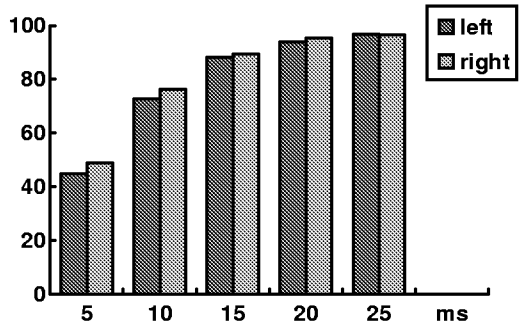

Fig. 2 - Boundary placement accuracy in automatic diphone segmentation.

The percentage of rule typology application, evaluated on the diphone decomposition of the entire database of about 700 sentences (more than 26000 diphones) is reported in Tab. 2.

Rule typology				
=	&	<, >	left boundary	right boundary
D	-	-	78.2 %	83.5 %
D	&	E	1.0 %	0.9%
E	&	D	7.2 %	5.0 %
SVF	&	D	13.4 %	10.4 %
SVF	&	E	0.2 %	0.2 %

Tab. 2 - Statistics of rule typology application.

Duration rules are the most frequently applied, because of their generality, direct evidence and widespread use in many contexts, especially for unvoiced sounds. Actually, during manual segmentation also expert phoneticians take advantage of the time structure as the main cue in many phonetic contexts. However, a significative amount of cases was better solved by jointly using energy and duration constraints. In this way, a compensation effect can be obtained for phone segmentation inaccuracy typical of some specific contexts. As an example, a local energy fall may be useful to determine a reliable range where boundaries of /r/ and /r:/ can be placed: in these cases it may be convenient to introduce different duration constraints, according to the preceding or the following contexts (vowels, consonants, "schwa", etc). Similarly, a joint use of energy, spectral variation cues and duration constraints was often applied to transitions including steady-state speech segments (e.g. vowels).
The diphone boundary accuracy may change according to the context: a decreasing performance was observed, in this order, for unvoiced stops, unvoiced fricatives and affricates, low energy voiced stops and fricatives, interconsonantal vowels, nasals and liquids, vowel sequences.

The system has been extensively applied for diphone segmentation of a very large database of more than 20000 isolated words (names, surnames, addresses, localities). About 24000 compound non-uniform units were derived to improve the quality and naturalness of an automated reverse telephone directory service.

7. CONCLUSION

A method for automatic speech unit extraction has been discussed. Statistical techniques for phone segmentation have been merged with acoustic/phonetic knowledge to bootstrap a rule-based system for diphone segmentation. Preliminary experiments, using the system in a speaker dependent mode, demonstrated a satisfactory performance that may be compared to that of an expert phonetician. Thanks to the system application to all the diphones of each utterance, the use of this system will favour the expansion of unit inventories. Another advantage is represented by the easy formulation and extension of rule sets. A drawback to address is represented by the dependence of system accuracy on the phonetic aligner performance: this limitation is more evident when the system operates in speaker independent mode and it may be overcome by improving the phone-unit modeling. Furthermore, next work will address the introduction of segmentation rules that may involve large phonetic contexts (e.g. left and right context-dependent or syllable-dependent rules). Finally, the introduction of a robust automatic F0 estimator as well as of a pitch marker assignment module is envisaged to make the system operating in pitch-synchronous fashion.

REFERENCES

[1] O. Boeffard, L. Miclet and S. White, *"Automatic generation of optimized unit dictionaries for text-to-speech synthesis"*, Proc. EUROSPEECH '93, Vol. 2, pp. 1211-1214, Berlin, 1993.

[2] A. Ljolje, J. Hirschberg and J. van Santen, *"Automatic speech segmentation for concatenative inventory selection"*, Proc. ESCA/IEEE Workshop on Speech Synthesis, pp. 93-96, New Paltz, September 1994.

[3] F. Brugnara, D. Falavigna and M. Omologo, "Automatic segmentation and labeling of speech based on Hidden Markov Models", *Speech Communication*, Vol. 12, no. 4, pp. 357-370, August 1993.

[4] B. Angelini, F. Brugnara, D. Falavigna, D. Giuliani, R. Gretter and M. Omologo, *"Automatic segmentation and labeling of English and Italian speech databases"*, Proc. EUROSPEECH '93, Vol. 1, pp. 653-656, Berlin, 1993.

[5] M. Balestri, S. Lazzaretto, P.L. Salza and S. Sandri, *"The CSELT system for Italian text-to-speech synthesis"*, Proc. EUROSPEECH '93, Vol. 3, pp. 2091-2094, Berlin, 1993.

[6] B. Angelini, F. Brugnara, D. Falavigna, D. Giuliani, R. Gretter and M. Omologo, *"A baseline of a speaker independent continuous speech recognizer of Italian"*, Proc. EUROSPEECH '93, Vol. 1, pp. 847-850, Berlin, 1993.

[7] P. Cosi, D. Falavigna and M. Omologo: *"A preliminary statistical evaluation of manual and automatic segmentation discrepancies"*, Proc. EUROSPEECH '91, Vol. 2, pp. 693-696, Genova, 1991.