QUALITY IMPROVEMENT OF (WIRELESS) PHONE-BASED TELE-SERVICES USING ADVANCED SPEECH SYNTHESIS TECHNIQUES

E. Keller

Laboratoire d'analyse informatique de la parole (LAIP), IMM-Lettres,
University of Lausanne, 1015 Lausanne, Switzerland
Phone: +41 21 692 3024; Fax: +41 21 692 3045
Email: eric.keller@imm.unil.ch

ABSTRACT

Keywords: ***speech synthesis, signal generation, prosody, sound quality*.** Speech transmission by (cellular) telephony encounters frequent difficulties in acoustically challenging circumstances, with auditorily handicapped populations, and with linguistically diverse target users. Traditionally, such transmission difficulties are multiplied when speech is generated by computer. Recent advances in speech synthesis technology have started to bridge these gaps by the development of advanced signal generation and prosody calculation models.

1. INTRODUCTION

Speech synthesis devices are increasingly integrated into everyday communication facilities. Telecommunication services are beginning to furnish directory information, airline and railroad arrival and departure times, stock quotations and product information directly from text. Computer manufacturers include speech synthesis into standard interfaces for email reading and to facilitate text access for the visually handicapped. Car radio and cell telephone producers are pursuing projects to integrate speech synthesis with traffic guidance systems. Finally, signal transmission engineers are experimenting with speech synthesis as an ultra-low bitrate transmission technique, especially for wireless and submarine applications. All of these applications pose extraordinary challenges, primarily with respect to transmission quality, and secondarily in terms of the computational footprint. In this presentation, I wish to outline the status and limits of typical current systems, and to characterise developments that can be expected over the next ten years.

It is important to clarify our terminology, and that in terms of the typical processing components of a text-to-speech system (TTS). Most TTS are structured as shown in Figure 1. A text is first translated into an *annotated phonetic representation*. Phonetic and annotation symbols are required in the next step to calculate a set of *prosodic parameters*. These parameters are subsequently used in the generation of a *speech signal waveform*. The three key components of a TTS are thus (a) the text-to-phoneme component, (b) the prosodic interpretation component, and (c) the signal generation component.

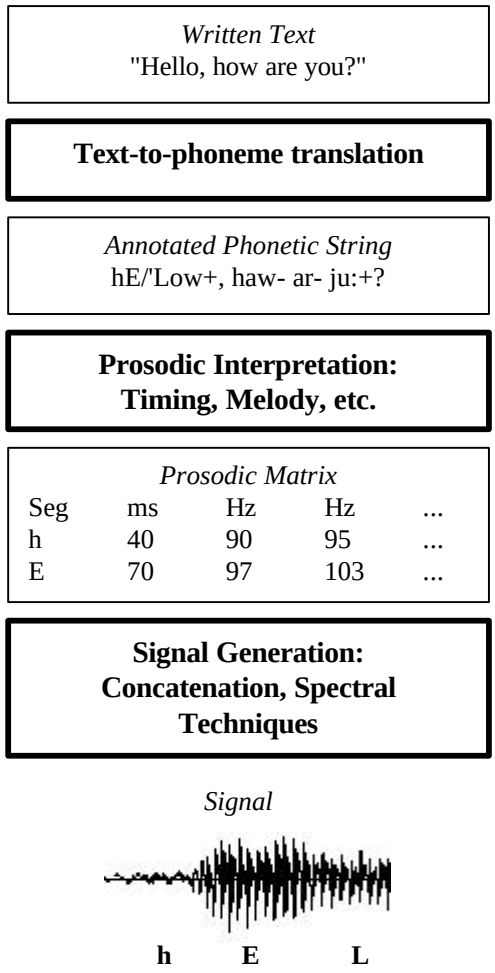

Figure 1: Typical architecture of a text-to-speech system. The light-bordered boxes show in- and outputs, and the heavy-bordered boxes show processing stages.

2. COMPONENTS: THE CLOSER TO THE SURFACE, THE MORE CRITICAL

Certain of these processing steps are particularly quality-critical. In general, "criticality" increases towards the end of the processing chain. The most critical component for high user acceptance is the *signal generation component*. Expressed in down-to-earth terms, quality at this level makes the difference between user acceptance and user rejection.

At the next higher level, inadequacies in *prosodic processing* can be very disturbing as well. For example, pauses in the wrong place, or intonational miscalculations by only 10 Hz, can introduce serious perceptual disruptions. If a system causes an excessive number of misunderstandings, or if a system's prosody is overly repetitive, user annoyance and user rejection is the result.

Errors at the highest, the *text-to-phonetics* level are, on the whole, the least critical. A certain number of phonetic errors are inevitable, given the impossibility of storing the correct pronunciation of every single word of a language, of every geographic locality, of all of the world's brand names, or of every person's proper name. Typical French TTS have been found make a phonetic error every 30 to 200 words [1]. Although systems that make fewer errors are unquestionably more useful than systems that make more errors, a certain redundancy characterizes human language. This quite often works in favor of the acceptance of speech synthesis systems despite errors at this level. (Obviously, a certain class of messages, such as names and numbers from a telephone directory, is excluded from this advantage, since it does not benefit from much redundancy.)

As a consequence of this "bottom-first" principle, much effort is currently being channeled into the improvement of models in signal generation and prosodic processing. In fact, the COST 258 European research project (1996-2000) which I have the honour of presiding, is specifically concerned with identifying and pursuing vectors of improvement in these two domains of TTS development[1]. Some interesting directions are being delineated in this project, which I shall turn to next[2].

3. SIGNAL GENERATION: TO GO BEYOND CONCATENATION

Most observers of information technology recall the robot-like sound quality of the first fully operational TTS, such as DEC-Talk for English. By contrast, more recent systems show much better quality (for examples, please consult the web pages of our COST 258 member laboratories). The interesting point about this historical development was that it was by and large due to an increase in the detail and sophistication of the underlying speech signal model. The rather poor quality of speech synthesis systems produced before 1990 was essentially due to an inadequate modelling of the *high-frequency* and *segment-transitional* properties of the speech signal [2]. Current signal generation models satisfy much better the full panoply of requirements at these two levels.

But as good as the best of our current systems are, they are still rather limited. Most of today's commercial and successful laboratory systems are *concatenative systems*, i.e., they are based upon a process of chaining prerecorded speech signal portions (typically diphones, *i.e.*, transitions between two phones or speech sounds), and of modifying their fundamental frequency and timing structure. Such systems have the inherent *advantage* of capturing all the acoustic complexity of a real speaker (including high-frequency components and the segment-transition information). But at the same time, they are also subject to some intrinsic limitations.

The most notorious limitation of such a system is that each new voice and each new speech style requires the rerecording, resegmentation and reanalysis of a complete new diphone inventory. This requirement is not to be underestimated. Even if new procedures can derive diphone databases essentially automatically from a set of recordings, the requirement for obtaining a new recording for each new style of speech or each new voice remains. In concrete terms, if one wishes a first, complete set of diphones for clear, deliberate speech of a language like English (2200 diphones), this requires a day's recording with a good, professional speaker. If subsequently, one wishes fast speech, that's another day's recording. *A whole new set of recordings* is required for each new voice, such as male, female, children's or age-marked voices. Further multiplications are required for speech with different emotions, such as depressive-tone speech, angry speech, sad speech, etc. And in the end, the developer cannot even be sure that a sentence composed of diphones recorded on different days will

[1] For more information about COST 258, please consult **http://www.unil.ch/imm/docs/LAIP/COST_258/cost258.htm**.

[2] These are directions as seen from the chair, but since COST 258 is a very diverse group, it stands to reason that not every participant in the group shares the views expressed here.

sound natural. Speakers cannot easily maintain the same richness of expression from one recording session to the next, but introduce subtle, but auditorily relevant variations into the signal.

The complexity does not end there. Diphone transitions are not necessarily the same for all speech styles. In French, for example, the slow pronunciation of "semaine" has two syllables ("se-maine"), and the fast pronunciation has one syllable ("s'maine"). Although this suggests that in a concatenative approach, one could simply borrow the diphone [sm] taken from a loan word like "smog" for the beginning of the fast pronunciation, unfortunately that is not so. Careful phonetic analysis has shown that the [sm] in "smog" does not sound like the [sm] in a fast pronunciation of "semaine". If used in this manner, it would contribute to lack of naturalness in synthesis [3, 4].

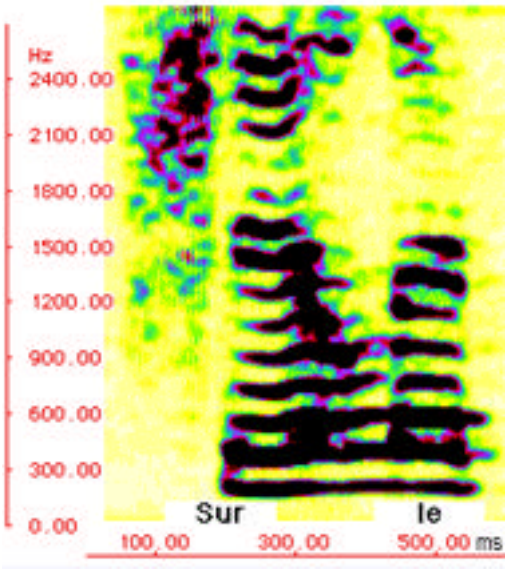

Figure 2: Spectrogram of a segment of speech showing the harmonic structure as well as high-frequency noise components. Female voice saying "sur le piano".

Efficiency is another issue that is barely being adressed by current concatenative signal generation techniques. An uncompressed concatenative database occupies a minimum of between 5 and 12 Mb of memory per style of speech[1], with each style of speech multiplying this base requirement. But much as in the MPEG3/4 standards for audio and video compression, only those aspects of the signal need to be preserved that actually contribute to clarity and distinctiveness. If the field evolves away from concatenative techniques towards more sophisticated signal generation techniques, it may be useful to consider algorithms that generate primarily perceptually efficient signal output.

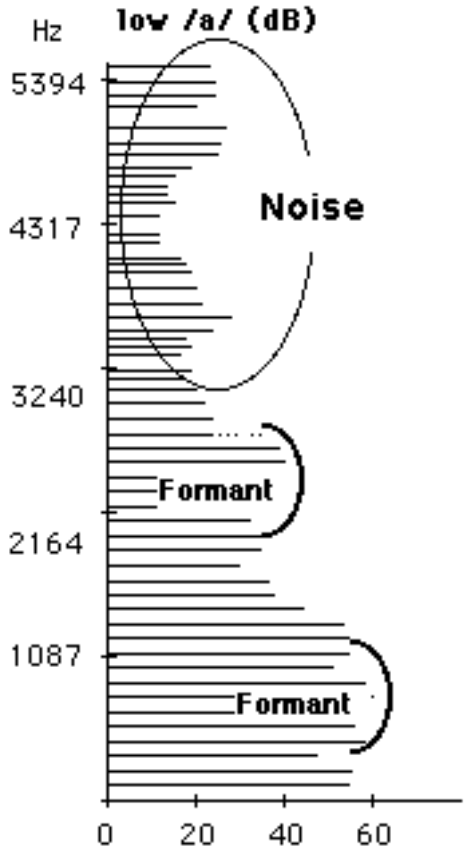

Figure 3: Harmonics extraction for a central portion of the sound [a]. It can be seen that lower harmonics are more regular than those in the higher spectral domain. This, in addition to high-frequency noise components of fricative and other unvoiced sounds, motivates separate processing for high-frequency components of the speech.

Taken together, these reflections permit to extrapolate that signal generation efforts of the next ten years will be concentrated on algorithms that satisfy the following three classes of requirements:

(a) *natural (contextually integrated) segment-to-segment transitions:* articulatorily and acoustically integrated segment-to-segment transitions for all desired speaking and speaker conditions (including *slow*, *fast*, and *interactive speech*, *male, female and children's speech*, etc.),

(b) *full prosodic modifiability:* this means, (1) prosody parameter ranges reflecting the full human range (concatenative systems place severe limitations upon F0 and intensity manipulation), and (2) the availability of a more extensive set of parameters, *i.e.*, not only *fundamental frequency* and *duration*, but also minimally *intensity* and *spectral tilt*, and perhaps other parameters.

[1] Depending on language. Languages like Italian and French can be represented with between 1200 and 1400 diphones, while languages like English and German require between 2200 and 2500 diphones [5].

(c) *efficiency*: primarily, *perceptually relevant aspects* of the signal are generated, and compression does not degrade the signal below worst-case requirements.

Only a small class of signal generation techniques satisfies all these requirements. Without striving for exhaustiveness, it would seem that the class of the sinusoidal-additive models stand an excellent chance of satisfying the given criteria. A prominent subclass of this group is known as Harmonics-and-Noise models [HNM]. Much work has been performed in this perspective recently [6-9].

In short, a given diphone is represented as the time-harmonics spectral amplitude matrix for the stretch of speech corresponding to the diphone (Figs. 2 and 3). In an initial approach, the signal is decomposed into its harmonics and noise components, and is recombined, after frequency and time modulation, into a contextually adequate new signal. In this approach to HNM modelling, spectral tilt or intensity are not manipulated. The main advantage over the concatenative approach lies in the facilitation of transition smoothing on the one hand, and in the greater ease and range of manipulation of fundamental frequency on the other.

In a more sophisticated variant of the HNM technique, the harmonics and noise matrix becomes the springboard for more extensive manipulations. Since the spectral placement of lower harmonics is quite regular, and since there is a predictable relationship between type of vowel and formant frequencies and amplitudes, matrix amplitudes in the lower frequency portion of the matrix can be summarized by functions. The higher harmonics of vowels are irregular, both with respect to harmonics placement and amplitude, and are thus best handled separately, either by models for higher-frequency behavior, or by explicit signal storage. Only this second, more sophisticated approach permits the additional manipulations of intensity and spectral tilt, required for handling differences in emotional state, semantic emphasis and age- and sex-characteristics of the voice. These higher-level functions can thus conceivably be used to generate different voices and expressive attitudes.

These new techniques come not without a cost. Beyond the not inconsiderable difficulty of rendering a full and clean HNM decomposition-recomposition framework operational, the main challenges lie in the definition of a complete set of predictive functions to describe speech in its full contextual and amplitude-nuanced expression. Much inspiration will no doubt come from traditional research into such factors (e.g. [10]), but there is no escaping that much new thinking will be required as well.

4. PROSODY: FROM LANGUAGE TO SPEAKER

In a series of publications, we have documented the severe limits of traditional theoretical assumptions concerning the generation of prosodic parameter specifications in speech synthesis ([11]-[13]). In short, traditional models of prosody have tried to capture the typical prosodic behavior of a given *language*. But speech synthesis must go beyond the lowest common denominators incorporated into language-based models. They require models of how a *specific speaker* of a language uses a given prosodic parametrisation in a *specific set of communicative circumstances*. Specifically,

(a) prosodic parameters for intonational contours, timing, as well as voice, emphasis and emotional attitude, should reflect the *behavior of a representative individual speaker of the language*,
(b) prosodic parameters should be distinguished for the *main styles of speech*, such as continuous, declarative text, the recitation of lists (*e.g.*, currency or stock market quotations), structured information (*e.g.*, address and phone number), and interactive speech,
(c) prosodic parameters should be *sufficiently nuanced* to provide satisfactory output when rendered by improved signal generation techniques.

By and large, current TTS do not yet satisfy these requirements. They still tend to model the *reading aloud* of continuous declarative text, inspired by traditional attempts to create "reading machines" [14]. But as has been indicated, prosody extends over much more than the reading of complete sentences, it includes a wide range of speech styles.

Furthermore, much effort has been characterized by the *"creation of grammars"*: grammars to relate syntactic structures to prosodic parameters [15] and grammars to define acceptable intonation contours [16]. But prosody expresses much more than syntactic states, it is for example prominently responsible for the expression of emphasis. Also, some of these grammars are also excessively simplistic (see Figs. 3 and 4). An excessively schematic, symbolic approach to prosodic encoding cannot hope to capture all the fine nuances that are required for the expression of emotional attitudes. Schematic and symbolic approaches will have to be

replaced by more sophisticated and numerically refined models.

Figure 3: Typical intonational contour.

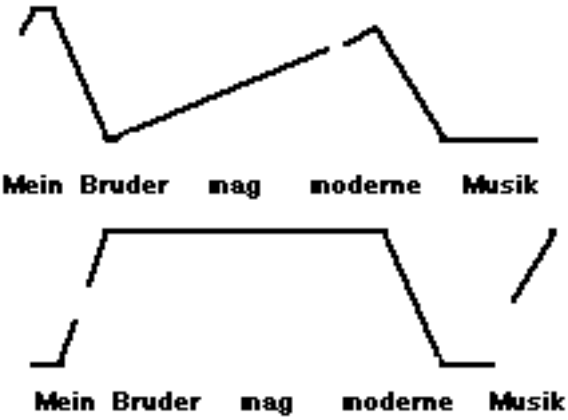

Figure 4: Traditional modelling of intonational contours. Two readings of the German sentence "Mein Bruder mag moderne Musik", adapted from an illustration appearing in 1998 on the web page of a prominent German speech synthesis laboratory. Even if a mediocre speech synthesis system gives a more or less acceptable rendering, a more sophisticated signal generation device will show that the long flat portion of the lower sentence is totally unnatural.

Furthermore, the output of some such grammars has been shown to be unrealistic. In our laboratory, putative relationships between major syntactic breaks and slowdowns and pauses in the utterance were not supported by actual measures of speech timing [17]. Rather, major breaks are related to chunking constraints, documented among others by the Neuchâtellois psycholinguist François Grosjean in an extensive series of studies [18, 19].

All this appears to have set into motion a veritable paradigmatic change with respect to prosodic prediction. As evidenced by recent international congresses, *automatic learning algorithms* such as neural networks are playing an important role in this change. This shift has been in preparation for some time. American and French laboratories report automatic F0 learning in 1994 [20, 21]. In Switzerland, the ETH-Laboratory in Zurich under the direction of Beat Pfister has used neural networks and statistical prediction for some time to derive models of fundamental frequency and timing [22-24]. More recently, a great number of European, American and Japanese laboratories have taken the same direction [25-27, etc.]. Automatic learning algorithms will clearly play a leading role in building more complete, more sophisticated and more precise models of prosodic parameter prediction.

This development is encouraging, but a note of caution is in order. Some practitioners of automatic learning have been heard to say that they wish to develop learning machines to derive entire prosodic models automatically for a given language. Given the great structural differences between languages, the great variety in prosodic expression, and in view of the great variety of speaking styles, such a hope appears quite unrealistic. Automatic learning algorithms will clearly accelerate the actual work to be done, and they will provide much more refined parameter predictions than grammar-based approaches. But the model constructor will still have to develop a clear theoretical understanding of the appropriate in- and outputs of the various components of the model. Only then can the lever be applied to the most appropriate corner of the object.

5. CONCLUSION

Traditional approaches to the two most critical components of speech synthesis processing, signal generation and prosodic processing, have recently been called into question. As speech synthesis requirements have expanded and quality requirements have become more stringent, concatenative approaches have come under pressure because of their extensive recording requirements and their inherent technical limitations. Sophisticated spectrally-based techniques are likely to take the relay, which will introduce new challenges for the definition and automatic generation of input parameter streams.

The requirements of greater naturalness have also exerted pressure upon traditional approaches to prosody modelling. Models originally developed to account for the generalities of a given language, or for the reading aloud of declarative texts, have been ill equipped to serve as models for the expression of a specific speaker in a variety of communicative situations. Predictions have often been crude or simply erroneous. Automatic learning algorithms are likely to play a major role in the development of more sophisticated and more nuanced models of the relationship between text input and prosodic output. However, the definition of input-output relations must be guided by a theoretical understanding of distinctive factors and language specificities.

6. ACKNOWLEDGEMENTS

Supported by the Swiss Office for Education and Science, project COST 258, by the EC under COST 258, and by Swiss federal KTI Project support.

7. REFERENCES

[1] Boula de Mareüil, P., Yvon, F., d'Alessandro, C., Aubergé, V., Bagein, M., Bailly, G., Béchet, F., Foukia, S., Goldman, J.-P., Keller, E., O'Shaughnessy, D., Pagel, V., Sannier, F., Véronis, J., & Zellner, B. (1998). Evaluation of grapheme-to-phoneme conversion for text-to-tpeech synthesis in French.*Proceedings of First International Conference on Language Resources & Evaluation* (pp. 641-645). Granada, Spain.

[2] Keller, E. (1997). Les théories de la parole dans l'éprouvette de la synthèse. In E. Keller, & B. Zellner (éds.). *Études des Lettres, vol 3.: Les défis actuels en synthèse de la parole.* (pp. 9-27). Lausanne: Université de Lausanne.

[3] Local, J. (1997). Ce qu'on peut faire pour la synthèse avec une meilleure prosodie et une meilleure qualité de signal. In Keller, E., & Zellner, B. (Eds.) 1997. *Études des Lettres, vol 3.: Les défis actuels en synthèse de la parole.* [pp. 29-46]. Université de Lausanne.

[4] Local, J. (1997). What some more prosody and signal quality can do for speech synthesis. (pp. 77-84). Proceedings of *Speech Technology in the Public Telephone Network: Where are we Today?* Rhodes, Greece. September 1997.

[5] Traber, C., Huber, K., Nedir, K., Jantzen, V., Keller, E., & Zellner, B. From multilingual to polyglot speech synthesis. In *Proceedings of Eurospeech 99*. Budapest (to appear).

[6] Stylianou, Y. (1996). *Harmonic plus Noise Models for Speech, combined with Statistical Methods, for Speech and Speaker Modification.* Ph.D. thesis, École nationale Supérieure des Télécommunications.

[7] Stylianou, Y., Dutoit, Th., & Schroeter, J. (1997). Diphone concatenation using a harmonic plus noise model of speech. *Proceedings of Eurospeech 97*, vol. 2, pp. 613-616. Rhodes, Greece.

[8] Bailly, B., Bernard, E., & Coisnon, P. (1998). *Sinusoidal modelling*. 4th COST 258 Meeting, Vigo, Spain.

[9] Banga, E.R., Salgado, X.R., Mateo, C.G. (1998). *Concatenative text-to-speech synthesis based on sinusoidal modelling.* 4th COST 258 Meeting, Vigo, Spain.

[10] Stevens, K.N., (1999). *Acoustic Phonetics*. Current Studies in Linguistics Series, No. 30.

[11] Zellner, B. (1996). Structures temporelles et structures prosodiques en français lu. *Revue Française de Linguistique Appliquée: La communication parlée. 1*. 7-23. Paris.

[12] Keller, E., Zellner, B., & Werner, S. (1997). Improvements in prosodic processing for speech synthesis. Proceedings of *Speech Technology in the Public Telephone Network: Where are we Today?* Rhodes, Greece. September 1997.

[13] Keller, E., & Zellner, B (1998). Motivations for the prosodic predictive chain. *Proceedings of ESCA Symposium on Speech Synthesis*. Paper 76, pp. 137-141. Jenolan Caves, Australia.

[14] Campbell, N. (1998). Where is the information in speech? (and to what extent can it be modelled in synthesis?) *Proceedings of the Third ESCA/COCOSDA Workshop on Speech Synthesis* (pp. 17-20). Jenolan Caves, Australia.

[15] Hirst & Di Cristo (eds), *Intonation Systems. A survey of twenty languages.* Cambridge, Cambridge University Press.

[16] t'Hart, J. Collier, R., & Cohen, A. (1990). *A Perceptual Study of Intonation.* Cambridge, UK: Cambridge University Press.

[17] Zellner, B. (1998). *Caractérisation et prédiction du débit de parole en français. Une étude de cas.* Thèse de Doctorat. Faculté des Lettres, Université de Lausanne.

[18] Grosjean, F. (1983). How long is the sentence? Prediction and prosody in the on-line processing of language. Linguistics, 21. 501-529.

[19] Monnin, P & Grosjean, F. (1993). Les structures de performance en français: caractérisation et prédiction. *L'Année Psychologique*, 93, 9-30.

[20] Rahim, G.M. (1994). *Artificial Neural Networks for Speech Analysis/Synthesis.* Chapman & Hall.

[21] Emerard, F., Mortamet, L., & Cozannet, A. (1994). Prosodic processing in a text-to-speech synthesis system using a database and learning procedures. In G. Bailly & C. Benoit (eds.), *Talking Machines* (pp. 225-254). North-Holland.

[22] Huber, K. (1991). *Messung und Modellierung der Segmentdauer für die Synthese deutcher Lautsprache.* Diss. ETH Nr. 9535.

[23] Traber, C. (1992). F0 generation with a database of natural F0 patterns and with a neural network. In G. Bailly & C. Benoit (eds.), *Talking Machines* (pp. 287-304). North-Holland.

[24] Riedi, M.P. (1998). *Controlling Segmental Duration in Speech Synthesis Systems.* Diss. ETH No. 12487 & TIK-Schriftenreihe Nr. 26.

[25] Keller, E., & Zellner, B. (1996). A timing model for fast French. *York Papers in Linguistics*, 17, University of York. 53-75.

[26] Bagshaw, P.C. (1998). Unsupervised training of phone duration and energy models for text-to-speech synthesis. *Proceedings of the 5th International Conference on Spoken Language Processing* (paper Tu3A1), pp. 1-133. Sydney, Australia.

[27] Streefkerk, B.M., & Pols, L.C.W. (1998). Automatic detection of prominence (as defined by listeners'

judgements) in read aloud Dutch sentences. *Proceedings of the 5th International Conference on Spoken Language Processing* (paper Tu5P34), pp. 1-173. Sydney, Australia.